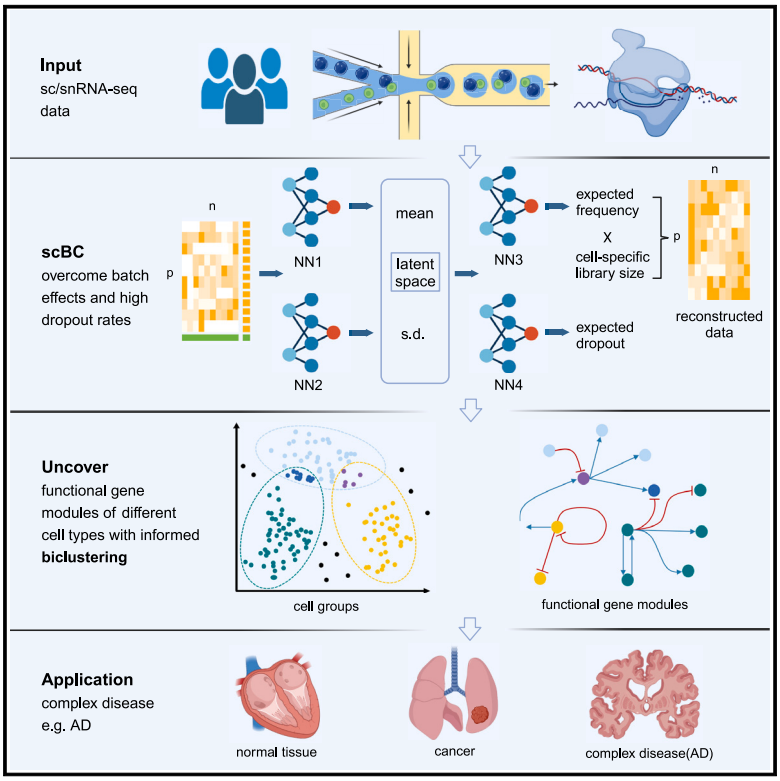
Cell Reports
Methods

Article

Single-cell biclustering for cell-specific transcriptomic perturbation detection in AD progression

Graphical abstract



Highlights

- scBC detects gene network biomarkers in scRNA and snRNA-seq data
- scBC incorporates existing biological information to guide single-cell biclustering
- scBC outperforms other biclustering methods in a variety of settings
- scBC reveals cell-specific gene module perturbations in Alzheimer disease

Authors

Yuqiao Gong, Jingsi Xu, Maoying Wu, Ruitian Gao, Jianle Sun, Zhangsheng Yu, Yue Zhang

Correspondence

yuzhangsheng@sjtu.edu.cn (Z.Y.), yue.zhang@sjtu.edu.cn (Y.Z.)

In brief

Gong et al. develop a single-cell Bayesian biclustering (scBC) framework that uncovers functional gene-module perturbations of different cell types in Alzheimer disease. Using scRNA and snRNA-seq data, scBC detects gene network biomarkers, overcoming challenges such as batch effects and high dropout rates. Outperforming other methods, scBC provides insights into complex disease mechanisms.

Gong et al., 2024, Cell Reports Methods *4*, 100742
April 22, 2024
https://doi.org/10.1016/j.crmeth.2024.100742

CellPress

Please cite this article in press as: Gong et al., Single-cell biclustering for cell-specific transcriptomic perturbation detection in AD progression, Cell Reports Methods (2024), https://doi.org/10.1016/j.crmeth.2024.100742

Cell Reports Methods

CellPress
OPEN ACCESS

Article

Single-cell biclustering for cell-specific transcriptomic perturbation detection in AD progression

Yuqiao Gong,[1] Jingsi Xu,[1] Maoying Wu,[1] Ruitian Gao,[1] Jianle Sun,[1] Zhangsheng Yu,[1,2,3,4,*] and Yue Zhang[1,2,4,5,*]
[1]Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, 800 Dongchuan Road, Minhang District, Shanghai 200240, China
[2]SJTU-Yale Joint Center for Biostatistics and Data Science Organization, Shanghai Jiao Tong University, Shanghai, China
[3]Clinical Research Institute, Shanghai Jiao Tong University School of Medicine, Shanghai, China
[4]Center for Biomedical Data Science, Translational Science Institute, Shanghai Jiao Tong University School of Medicine, Shanghai, China
[5]Lead contact
*Correspondence: yuzhangsheng@sjtu.edu.cn (Z.Y.), yue.zhang@sjtu.edu.cn (Y.Z.)
https://doi.org/10.1016/j.crmeth.2024.100742

MOTIVATION Alzheimer disease (AD) is a highly complex and debilitating neurodegenerative disorder that has been the subject of extensive research and public attention in recent years. The pathogenesis of AD involves intricate changes to gene networks occurring across multiple cell types. Investigating individual genes or focusing on single cell types alone may therefore present limitations to fully comprehending the disease. We sought to develop a more comprehensive approach to analyze AD, one that can simultaneously capture the complexity of gene interactions and cellular heterogeneity.

SUMMARY

The pathogenesis of Alzheimer disease (AD) involves complex gene regulatory changes across different cell types. To help decipher this complexity, we introduce single-cell Bayesian biclustering (scBC), a framework for identifying cell-specific gene network biomarkers in scRNA and snRNA-seq data. Through biclustering, scBC enables the analysis of perturbations in functional gene modules at the single-cell level. Applying the scBC framework to AD snRNA-seq data reveals the perturbations within gene modules across distinct cell groups and sheds light on gene-cell correlations during AD progression. Notably, our method helps to overcome common challenges in single-cell data analysis, including batch effects and dropout events. Incorporating prior knowledge further enables the framework to yield more biologically interpretable results. Comparative analyses on simulated and real-world datasets demonstrate the precision and robustness of our approach compared to other state-of-the-art biclustering methods. scBC holds potential for unraveling the mechanisms underlying polygenic diseases characterized by intricate gene coexpression patterns.

INTRODUCTION

In recent years, the advancement of single-cell sequencing technology has enabled the analysis of single-cell data to reveal meaningful biological information at the cellular level. Specifically, single-cell RNA sequencing (scRNA-seq) enables the sequencing of cells that are hard to retrieve or challenging to isolate.[1] This unprecedented resolution into cell states provides us with new insights into the function and dysfunction of cells,[2] which is particularly necessary for complex diseases such as Alzheimer disease (AD), because changes in gene expression are related to cell type.[3,4] Recently, there has been a surge of single-cell studies aimed at understanding the mechanism of AD based on transcriptional profiles,[3,5,6] which have provided valuable insights into cellular diversity. However, these studies often lack an integrative analysis of functional gene modules (FGMs), which can reveal how genes work together to regulate biological processes. A recent study used a network-based approach to identify FGMs involved in the selective vulnerability of neurons in AD, demonstrating the importance of analyzing FGMs to gain insights into the underlying mechanisms of complex diseases such as AD.[7] FGMs are groups of genes that work together to perform a specific biological function and can exhibit complex coexpression or co-regulation patterns, rather than solely comprising differentially expressed genes.[8,9] Moreover, these local patterns are often cell specific and may change with disease progression.[10–12] Therefore, it is crucial to identify FGMs and their corresponding functional cell groups

simultaneously in studies of complex diseases. In this study, we focus on FGMs as gene network biomarkers to investigate their potential role in AD.

Unlike clustering methods, which can only conduct clustering in either cell space or gene space, biclustering can identify FGMs and their corresponding functional cell groups simultaneously. A cells-genes pair is called a bicluster, and the genes in a specific bicluster can be deemed as an FGM shared across related cells. Therefore, through biclustering, we can easily identify FGMs and find the cell populations in which they are active at the same time. It is worth noting that in the complex cell machinery, multiple FGMs are active in a cell group, and different cell groups may share a common FGM (Figure S1). Fortunately, biclustering with overlapped biclusters can easily capture such complex features.[13] Through biclustering, cell population-specific gene network biomarkers and potential gene-cell connections can be identified in a single pass.

Although biclustering is an exquisite tool, it encounters some problems when applied to scRNA-seq data. First, batch effects, due to laboratory conditions, reagent lots, and personnel differences, are widespread and critical to address.[14] If batch effects are not properly accounted for, then biclustering algorithms may falsely identify batch-specific coexpression patterns instead of true biological patterns, leading to incorrect conclusions. Second, due to low mRNA content per cell and molecule losses during the experiment (known as "dropout"), the gene expression matrix has a substantial amount of zero read counts that can cause problems for biclustering algorithms that assume continuous expression values.[15] Biclustering algorithms that are not designed to handle dropout may either ignore the zero read counts, leading to incomplete biclusters, or consider them to be low-expressed genes, leading to spurious biclusters. Furthermore, the selection of a specific scRNA-seq protocol (e.g., droplet-based methods such as 10X Genomics, plate-based methods such as Smart-seq2) can significantly influence both the magnitude and characteristics of dropout events. Moreover, variations in sequencing depth can introduce variability in the detection limit of low-abundance transcripts, thereby resulting in diverse levels of dropout occurrences. Consequently, the development of a biclustering method that is adaptable to different sequencing protocols, accounting for their distinct dropout effects, would render it more widely applicable and relevant for comprehensive analysis of scRNA-seq data.

Although algorithms have been designed to address these inherent problems pervasive in scRNA-seq data, they have typically focused on improving the performance of the biclustering algorithm in one particular aspect—cell clustering, FGM finding, or the simultaneity of coclustering.[16–18] However, in complex polygenic diseases, functionally related potential cell groups are finely divided, cell-type conditional gene coexpression patterns are complicated, and the cell-gene correlation changes throughout the progression. Therefore, an algorithm with better performance in functionally related cell group discovery, FGM finding, and cell-gene correlation pattern detection would help to advance research into these diseases.

Here, we propose a single-cell Bayesian biclustering (scBC) method that can handle the problems mentioned above. We use a variational autoencoder (VAE) to model gene expression in single cells, enabling us to gracefully remove batch effects and impute missing data.[19] By estimating the variational posterior distribution, we can obtain a low-dimensional representation of each cell that is conditioned on the batch annotation, enabling us to obtain batch-corrected expression through the generating process. In addition, we can manually control the procedure of dropout during the generating process, leading to the predropout imputed expression. By reconstructing the original data matrix in this way, we can obtain more precise results when conducting biclustering. Furthermore, we incorporate existing biological information (e.g., gene interaction and regulation) into the biclustering procedure through the Bayesian framework, which guides variable selection to more likely capture pathway information and true biological signals.[20] The flowchart of our procedure is depicted in Figure 1.

RESULTS

scBC outperforms other methods on FGM detection

To investigate whether scBC can detect biologically meaningful FGMs, we analyzed four highly heterogeneous single-cell datasets obtained from different parts and tissues of the human body under different pathological conditions (purified peripheral blood mononuclear cell dataset, PBMC; cardiac cells with annotation from Heart Cell Atlas, HEART; scRNA-seq of lung adenocarcinoma, LUAD; and scRNA-seq of primary breast cancer. An outlook of these datasets can be found in Table S1). Figure 2A illustrates the preprocessing procedure, and the STAR Methods section provides further details. To verify the feasibility of our method, we compared scBC with six traditional biclustering algorithms, namely CC,[21] xMotif,[22] FABIA,[13] Bimax,[23] PLAID,[24] and GBC,[20] and two newly developed biclustering algorithms intended for scRNA-seq data: QUBIC2[17] and DivBiclust.[16] We also incorporated one brand new VAE architecture, autoCell,[25] into the data reconstruction procedure to see how scBC outperforms alternative choices. For methods that need to set the number of biclusters in advance (e.g., xMotif, CC, Bimax, FABIA, GBC, and scBC), we set the maximum number of biclusters as the number of cell types (based on cell label). We then conducted Gene Ontology (GO) enrichment analysis for each FGM detected by each method (see STAR Methods), and used $-\log10(p)$ (Benjamini-Hochberg [BH] adjusted) as the enrichment score. Methods that failed to detect any bicluster were assigned a score of zero (Figures 2B–2D). Since DivBiclust is a biclustering-based method for cell population discovery that only outputs the cell clustering result, it was overlooked in FGM identification comparisons.

We found that the FGMs detected by scBC were consistently more significant than those detected by other algorithms, even in highly heterogeneous settings (Figure 2B; Table S2). In the PBMC dataset, all of the methods were able to capture the specific FGM, indicating a relatively simple data structure. Among these methods, scBC performed the best, followed by CC. autoCell, GBC, xMotif, and FABIA also performed well, but PLAID, Bimax, and QUBIC2 gave unsatisfactory results. In the HEART dataset, both xMotif and Bimax failed to identify any biologically meaningful gene modules, whereas PLAID exhibited limited effectiveness (Figure 2B). These results suggest that biclustering

Please cite this article in press as: Gong et al., Single-cell biclustering for cell-specific transcriptomic perturbation detection in AD progression, Cell Reports Methods (2024), https://doi.org/10.1016/j.crmeth.2024.100742

CellPress
OPEN ACCESS

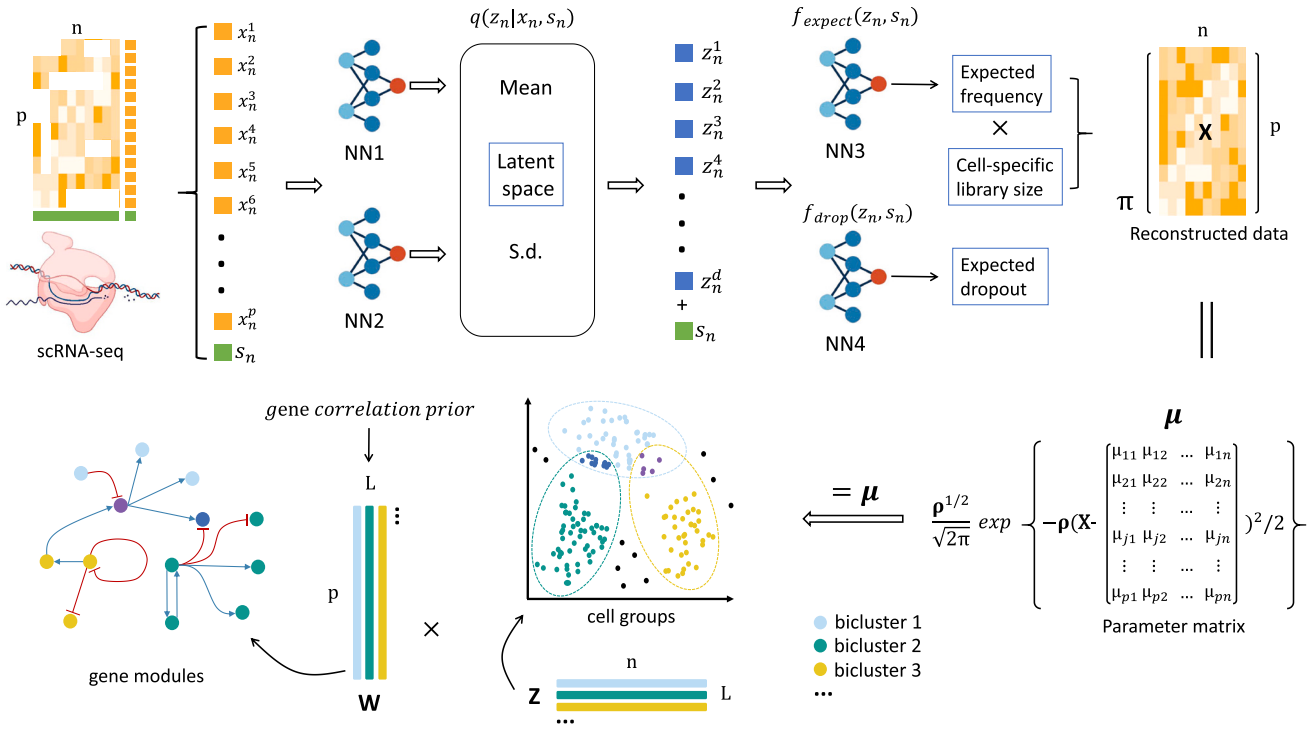


Figure 1. Flowchart of scBC procedure

The scRNA-seq data with high proportion of dropout and batch annotation (if available) is first fed into the VAE. We use x_n^g to denote the g_{th} gene in cell n. s_n is an extra dimension added for each cell to denote the batch annotation. Through the training process, we can get a low dimensional approximate posterior distribution $q(z_n|x_n, s_n)$ conditional on s_n. At the inference stage, the low dimensional representation of each cell z_n is taken to reconstruct the expression data through nonlinear mapping. The likelihood function of gene g from cell n is $\pi(x_{ji}|\rho_j,\mu_{ji}) = \frac{\rho_j^{1/2}}{\sqrt{2\pi}} exp\{ -\rho_j(x_{ji} - \mu_{ji})^2/2\}$. To reduce randomness, we decompose the parameter matrix $\boldsymbol{\mu}$ of the reconstructed data matrix **X** rather than directly on **X**. Gene correlation prior is used to guide the variable selection at the gene level, and results are presented as matrix **W**, with each column denoting a module. Matrix **Z** is the result at the cell level, with each row representing a functionally related cell group.

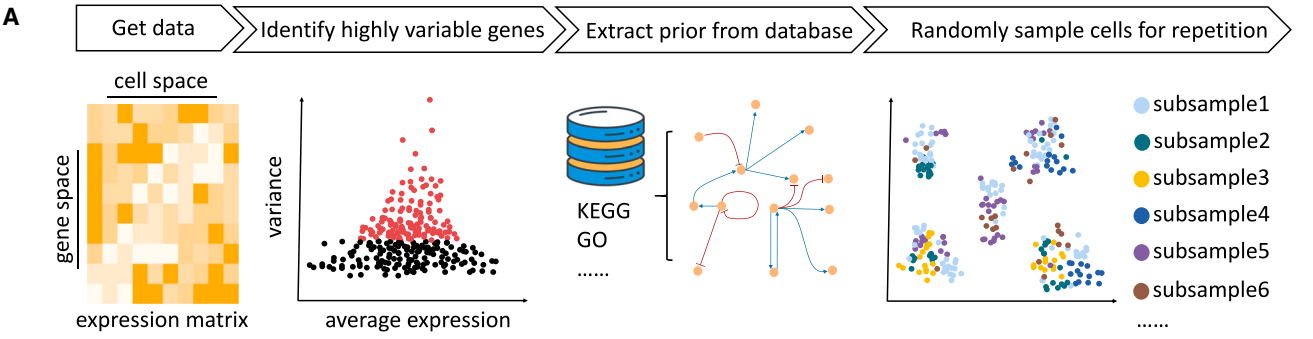
on cardiac tissue data presents greater challenges. However, GBC and CC demonstrated satisfactory performance, ranking second only to scBC and autoCell. Notably, scBC and autoCell exhibited remarkably similar and outstanding performance in this dataset (Figure 2B). However, in the LUAD dataset, which specifically pertains to the tumor-associated immune microenvironment, the performance of autoCell significantly deteriorated and lagged behind that of CC and GBC. This observation suggests a relatively weaker and less robust performance for autoCell in this context (Figure 2B). In the breast cancer dataset, where sample size was minimal (only up to 300 cells after negative sampling) and tumor cells were mixed with normal cells, many methods failed to detect meaningful FGMs (Figure 2B). Even under such challenging conditions, scBC was able to identify more biologically meaningful FGMs. Despite both CC and GBC demonstrating similar excellent performance in the first few conditions, CC was far inferior to GBC in the breast cancer dataset. Interestingly, despite being a biclustering method designed for FGM identification, QUBIC2 did not exhibit remarkable performance in these datasets, particularly in the first three datasets. This could be attributed to the ability of QUBIC2 to only detect biclusters of relatively small scale, indicating its suitability for analyzing smaller datasets and lack of versatility. Overall, scBC demonstrated robust superior performance in FGM detection across highly heterogeneous conditions.

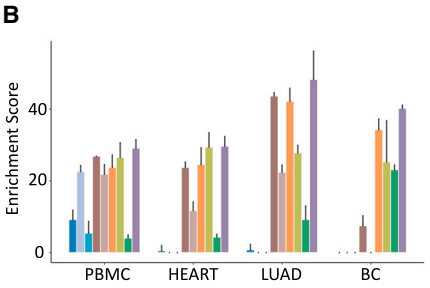
Since scBC is composed of several building blocks, we conducted an ablation study on all four highly heterogeneous datasets to investigate whether there are redundant components and determine which component contributed the most to the enhanced performance. In datasets of normal tissue, both the introduction of prior information and data reconstruction proved beneficial to performance. Furthermore, combining both strategies led to a more significant improvement (Figures S2A and S2B). For datasets of tumor-related tissue, it is intriguing to observe that a single strategy resulted in a detrimental performance, particularly the data reconstruction strategy. However, combining the two strategies led to an increased performance (Figures S2C and S2D). This finding suggests that due to the complexity of disease-related data, a single strategy is insufficient for these datasets. It also underscores the importance of combining the two strategies and highlights the potential of scBC in analyzing datasets related to complex diseases.

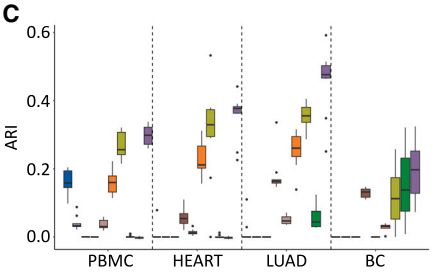
Benchmarking clustering results at cell level

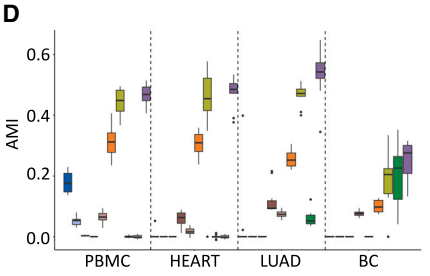
Intuitively, cell groups identified by biclustering are more functionally related since each group corresponds to a similar

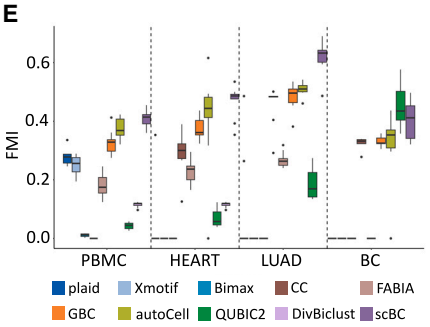
Please cite this article in press as: Gong et al., Single-cell biclustering for cell-specific transcriptomic perturbation detection in AD progression, Cell Reports Methods (2024), https://doi.org/10.1016/j.crmeth.2024.100742

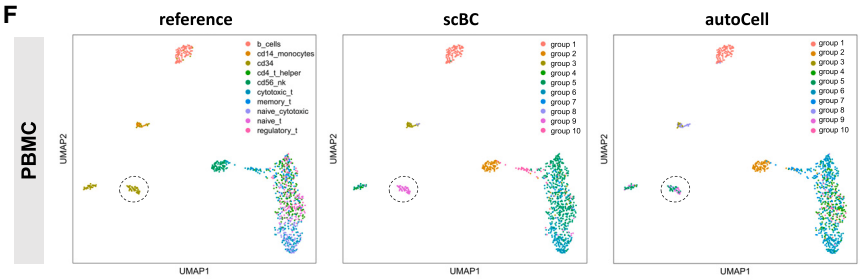


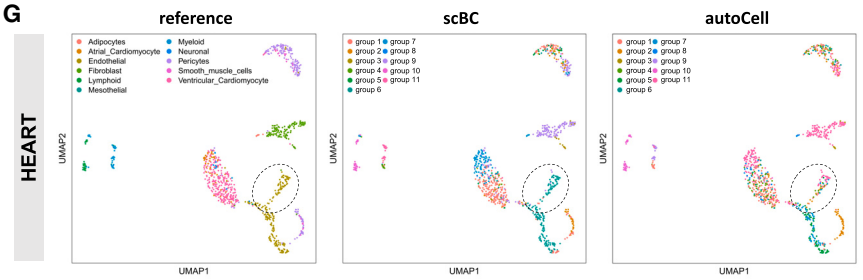


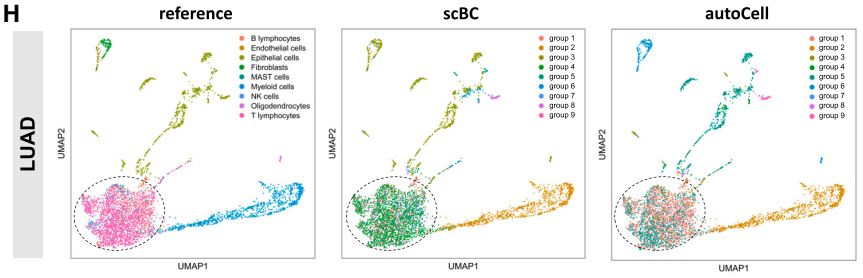


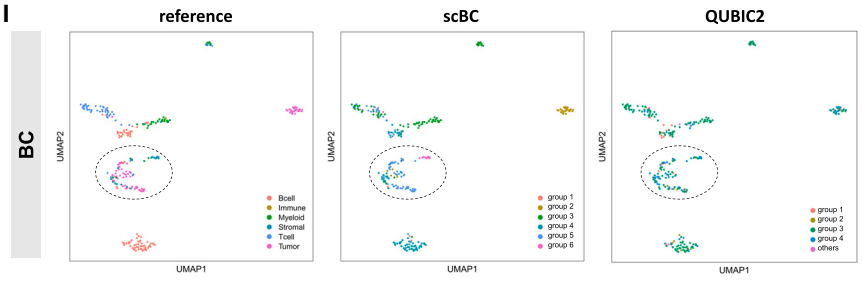












(legend on next page)

Please cite this article in press as: Gong et al., Single-cell biclustering for cell-specific transcriptomic perturbation detection in AD progression, Cell Reports Methods (2024), https://doi.org/10.1016/j.crmeth.2024.100742

FGM. Functionally related cells are also naturally more likely to belong to the same cell type since they have similar functions, although there can be exceptions. In this study, we investigated the clustering performance at the cell level to determine whether scBC provides more meaningful clustering results, even when focusing solely on the cell-level clustering results. As mentioned earlier, biclustering results may have overlap between each bicluster, which can result in single cells belonging to different groups. However, the results from scBC and GBC enable us to assign each cell to its most involved groups, which is also applied to the autoCell-based framework. For the remaining methods with less well-defined cell-level clustering results, we used the Markov clustering algorithm (MCL)[26] to transform the biclustering results, fully using the information from the biclustering results (see STAR Methods). We used the adjusted Rand index (ARI),[27,28] the Fowlkes-Mallows score (FMI),[29] and the Adjusted Mutual Information (AMI)[30] as recommended metrics to quantify the agreement between clusters (see STAR Methods). Their values range from −1 to 1, with higher values indicating better performance. We evaluated the clustering performance at the cell level using the four real-world datasets.

The cell clustering results obtained by scBC are consistently more precise than those of other methods across different heterogeneous conditions (Figures 2C–2E; Tables S3–S5). We found that Bimax performed the worst, not only in FGM detection but also cell clustering in all of the datasets. Although CC performed well in FGM detection in the PBMC dataset, it was unable to perform cell clustering tasks simultaneously. autoCell performed well in PBMC, HEART, and LUAD datasets, second only to scBC (Figures 2C–2E). In the HEART dataset, PLAID, xMotif, and Bimax were invalid (Figures 2C–2E), similar to the LUAD and breast cancer dataset. Notably, in the first three datasets, we observed that scBC was capable of capturing a pattern wherein certain cell populations consisted of a substantial number of a major cell type and a relatively smaller number of other cell types, whereas autoCell failed to do so (Figures 2F–2H). This indicates that scBC can accurately identify cell populations that comprise a combination of major and rare cell types. In the breast cancer dataset, QUBIC2 ranked second only to scBC in terms of ARI and AMI, but slightly outperformed scBC in terms of FMI (Figures 2C–2E), suggesting its suitability for analyzing small-scale data, which aligns with its performance in FGM detection. PLAID, xMotif, Bimax, and FABIA failed to cluster cells into functionally related groups because they cannot detect FGMs in the dataset. Unexpectedly, despite being a biclustering-based method intended for cell clustering, DivBiclust demonstrated limited potential in the PBMC and HEART datasets and even failed to identify cell clusters in the LUAD and breast cancer datasets (Figures 2C–2E). This could be attributed to the exceptionally high dropout rate in these datasets and the added complexity of noise in the tumor-related datasets. In a nutshell, these results demonstrate that scBC can capture the complex patterns involved in clustering functional cell groups and is more robust and precise than other methods across heterogeneous datasets.

scBC performs best on a bicluster level

Gene coexpression patterns differ across different cell types. These complex gene-cell correlations are of particular interest to us. When we compare the performance of different biclustering methods at the bicluster level, we pay more attention to whether the method can detect cell subgroups with similar FGMs and present these cells and FGMs at the same time. Once such a biosignal is found, we can make guidelines for downstream analysis. In this study, we introduced two evaluation methods, 1-CE and F score, to compare the performance of our scBC with other methods (see STAR Methods for simulation detail). Since DivBiclust only output cell-level results, it was not included in this benchmarking. We used simulated datasets with different dropouts under varying scales to elucidate how the performance of these methods varies along with the conditions (Figure 3; Tables S6–S11).

Since different scRNA-seq protocols often produce data with varying sparsity, our simulation started from dropout = 0.2 and explored with a step size of 0.1. When dropout was >0.5 we set step = 0.05 to get a more detailed performance variation in highly sparse cases. The results demonstrate that scBC consistently outperformed other methods in uncovering complex gene-cell correlation patterns, particularly with respect to F score, and the dropout rate was not excessively high (Figures 3C–3E, and 3G). FABIA and autoCell exhibited the second-best performance, with FABIA performing better with respect to F score (Figures 3B–3G). However, as the dataset size increased, the performance of FABIA became increasingly unstable (Figure 3F). PLAID showed an advantage in cases with minimal dropout, but its performance deteriorated rapidly as the dropout rate increased (Figures 3B–3D, and 3F). Conversely, methods such as CC, xMotif, and QUBIC2 were mostly ineffective across all of the settings (Figures 3B–3G). As expected, the ability of all of the methods to detect biclusters generally decreased with increasing dropout rate and dataset size (Figures 3B–3G). Nevertheless, scBC consistently outperformed other methods in the

Figure 2. scBC outperforms other methods in FGM detecting and cell clustering in 4 real-world datasets

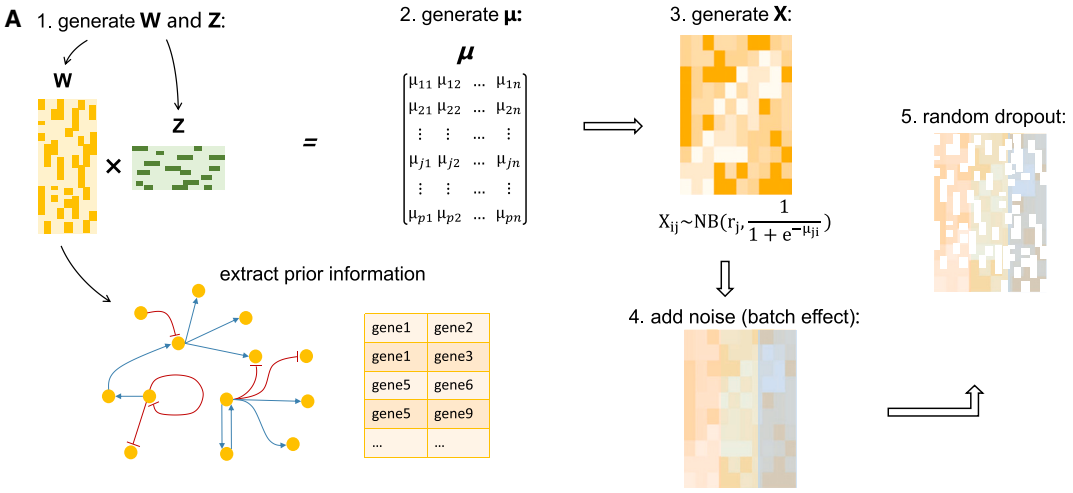
(A) The preprocess procedure of the 4 datasets. In each dataset, highly variable genes are filtered out and used to extract prior coexpression information in databases such as GO or the Kyoto Encyclopedia of Genes and Genomes. Cells are sampled along with the highly variable genes to generate 10 subsampled datasets for repetition.

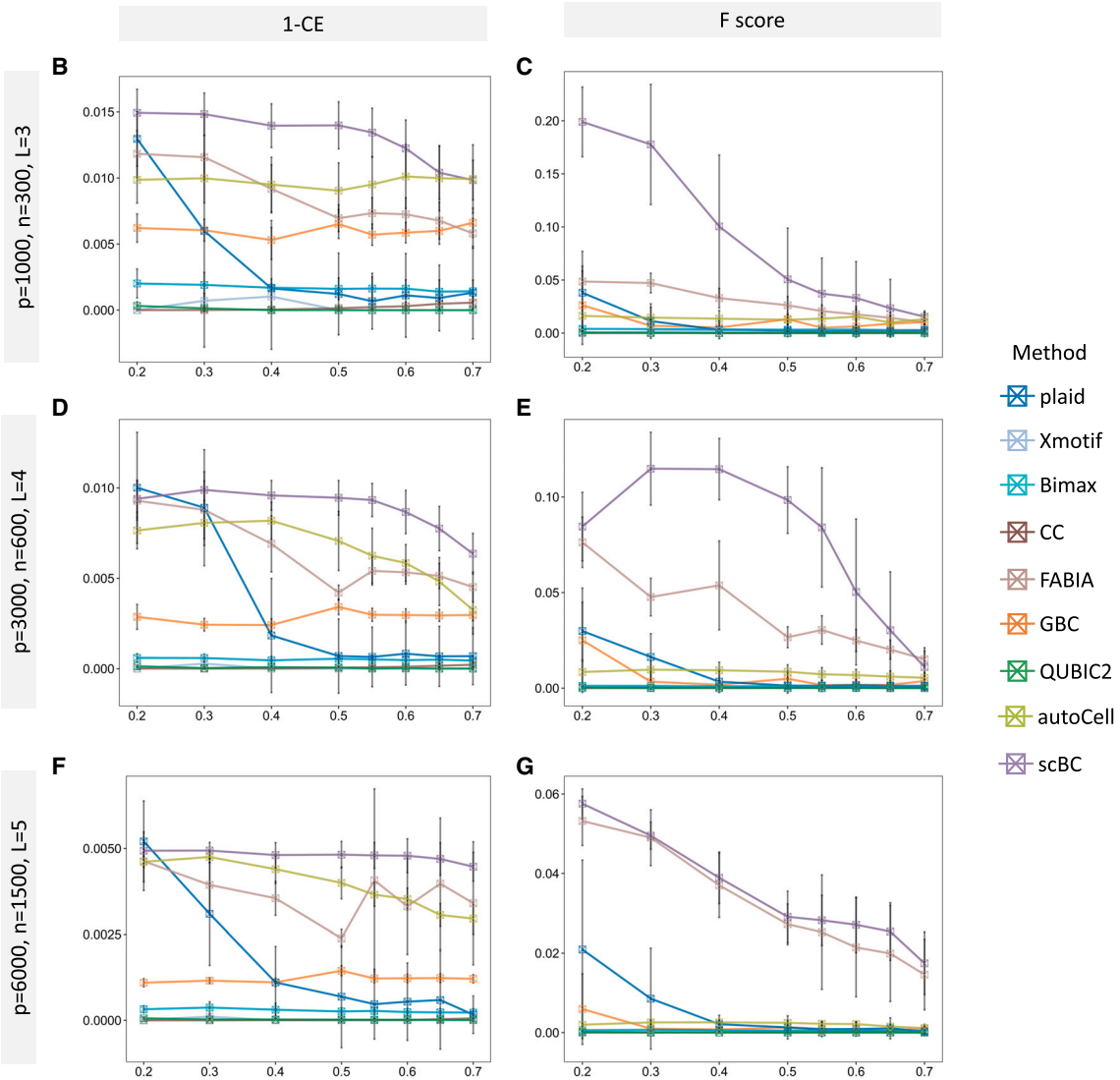
(B), Enrichment score of different methods in 4 highly heterogeneous datasets. The x axis represents different datasets and the y axis represents the enrichment score (−log10(p), BH adjusted) of different methods. The error bar stands for SD of the results of 10 subsamples.

(C–E) Benchmarking clustering results at cell level with ARI (C), AMI (D), and FMI (E) in the 4 datasets.

(F–I) Cell representation of UMAP dimensionality reduction. In addition to the reference labels, shown here are the methods with highest ARI (scBC) and second-highest ARI in the last subsample dataset (F, PBMC; G, HEART; H, LUAD; and I, breast cancer). The whole comparison of cell clustering can be found in Figures S3–S6. Some methods output too many categories, so that we merge some into "others" whose number of cells is <1% of the total sample. Highlighted with black dotted lines are cell populations' patterns that are correctly identified by scBC but not identified by the second-best method.

Please cite this article in press as: Gong et al., Single-cell biclustering for cell-specific transcriptomic perturbation detection in AD progression, Cell Reports Methods (2024), https://doi.org/10.1016/j.crmeth.2024.100742





(legend on next page)

Please cite this article in press as: Gong et al., Single-cell biclustering for cell-specific transcriptomic perturbation detection in AD progression, Cell Reports Methods (2024), https://doi.org/10.1016/j.crmeth.2024.100742

majority of cases, highlighting its superior reliability in capturing intricate gene-cell correlation patterns.

scBC uncovers the pathway perturbation in AD progression

Neuropsychiatric disorders involve complex polygenic determinants as well as brain alterations.[31] Biclustering methods can reveal cell population-specific gene coexpression patterns and discover potential gene-cell connections, making them inherently more suitable for the analysis and mining of complex polygenic disorders such as neurodegenerative diseases. At the same time, single-cell-level resolution is critical for neurodegenerative diseases such as AD because changes in gene expression are related to specific cell types.[3] Therefore, scBC is more reliable for analyzing the single-cell data of diseases with complex traits due to its excellent performance.

AD is a neurodegenerative disorder associated with aging, characterized by the accumulation of amyloid plaques and neurofibrillary tangles in the brain parenchyma. Recent research, using a single-nucleus RNA-seq (snRNA-seq) dataset from AD patients, has shown that AD is a complex disease involving multiple brain cell types, as evidenced by marker gene expression.[3] In the present study, we aim to investigate further transcriptomic perturbations during AD progression using gene network biomarkers identified by our scBC model. This dataset includes 48 postmortem human brain samples, with or without AD. The pathology groups are defined based on several pathological traits (Table S12): "no pathology" (no amyloid burden, no neurofibrillary tangles, and no cognitive impairment), "early pathology" (amyloid burden, but modest neurofibrillary tangles and modest cognitive impairment), and "late pathology" (higher amyloid burden, increased neurofibrillary tangles, global pathology, and cognitive impairment) (Figure 4A). After subsampling, we ensured that cells from different donors were well blended and not dominated by any one donor or biased by sex (Figures 4B–4D). We also ensured that cells of the same type across individuals were consistent (Figures 4E and 4F). To make sure that the results of multiple biclustering analyses corresponded with one another, we matched biclusters in different pathological progression stages and merged some FGMs according to the degree of overlap in gene sets (Figure 4G; STAR Methods). We found that the overlap of each FGM, as expected, is considerable (Figure 4H).

It is commonly believed that multiple FGMs can be simultaneously active in one cell type, and a single FGM can be shared across different cell types, but the composition percentage in different cells will vary. Our method, scBC, captured this structure perfectly (Figures 5A–5F), indicating that it is very suitable for such analysis. In this study, we focused on the perturbation of FGMs for each cell type during the progression of AD to gain a better understanding of the mechanisms underlying AD and to provide potential recommendations for therapy. To clarify the functional changes represented by specifically altered FGMs, we performed enrichment analysis of specific gene sets before and after a progression stage (see STAR Methods for details) to identify associated pathways that are disrupted during the progression (Figures 5G–5K). The complete enrichment analysis results can be found in Table S13.

For astrocytes and oligodendrocyte precursor cells (OPCs), FGM perturbation occurs almost only at early pathology (Figures 5A and 5F), indicating transcriptional patterns have largely changed in these two cell types before an individual develops severe pathological features, which is consistent with previous research.[3] Inhibitory neurons' change in FGM throughout the disease progression is minimal (Figure 5C), indicating that this cell type does not have many alterations in transcriptional patterns during AD progression.

Astrocytes are involved in neuronal trophic support, extracellular ion homeostasis, and brain fluid balance.[32] Energy metabolism is largely altered in AD astrocytes (Figure 5G), indicating the inflammatory state of the brain following injury and neurodegeneration since astrocytes are a central driver of energy homeostasis in the brain, which is also mentioned in previous studies.[32,33] Consistent with previous studies, we found that ion transporters are dysregulated in AD astrocytes (Figure 5G). At the same time, we also found that pathways related to myelination and neuron ensheathment are altered with the progression of AD (Figure 5G).

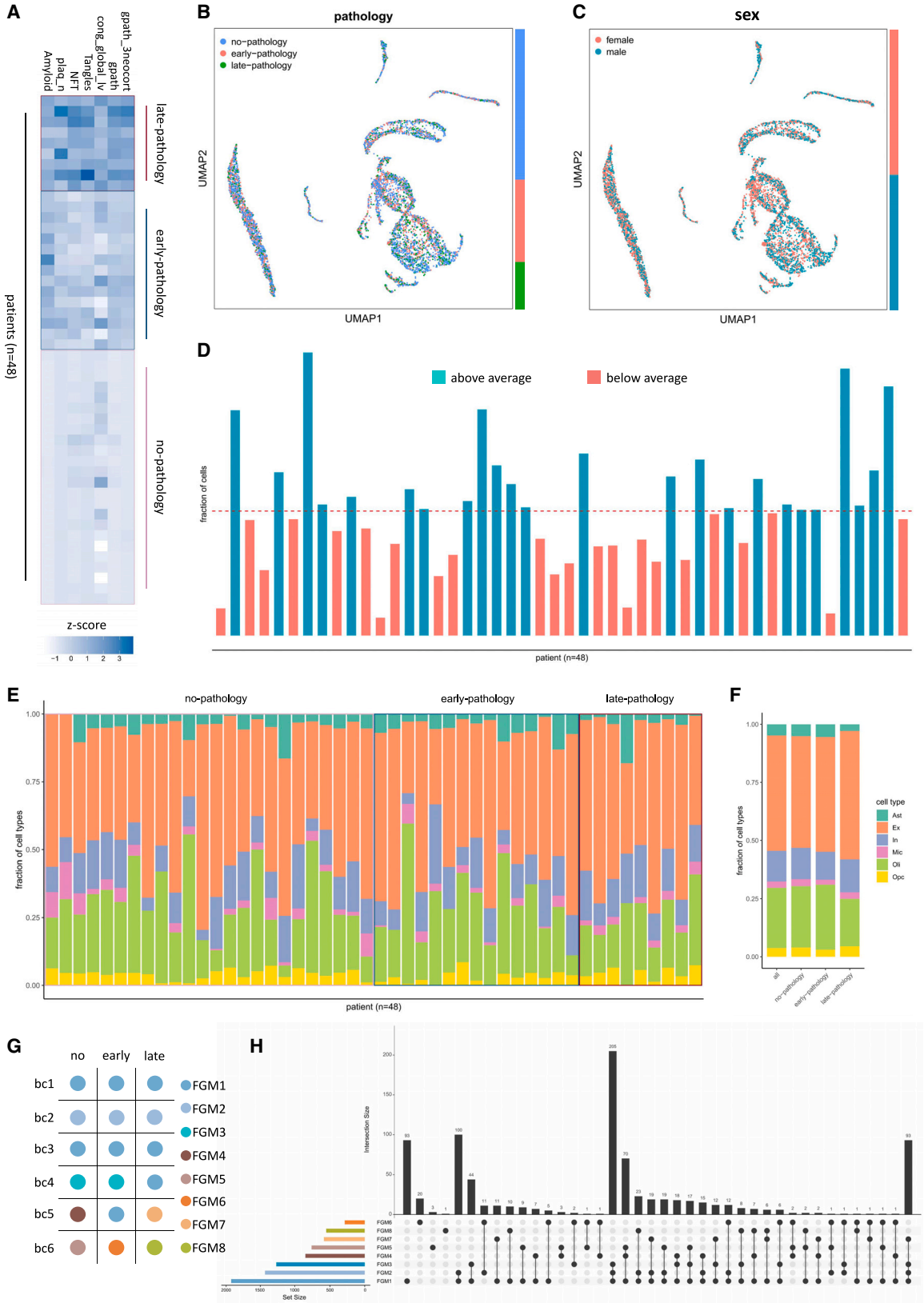
It has been found that gliogenesis and neuron ensheathment-related pathways are largely impaired in AD progression.[34] We found the same conclusion in the progression of AD pathology in excitatory neurons (Figure 5H). However, the transition of FGM composition in excitatory neurons from the normal state to the early-pathology state appears to be relatively subtle, in comparison to the significant change observed from the early stage to the late stage (Figure 5B). This suggests that the major perturbation of FGMs in this cell type primarily occurs during the late-pathology state. Another continuous change in excitatory neurons in AD progression is a general dysregulation in kinase activity (Figure 5H), which is closely related to neuronal DNA damage, well known to occur in AD neurons.[35] Previous studies mentioned that the immune response is also affected in the progression of AD.[32] Here, we identified a specific gene in late pathology, VSIG4 (see Table S13) that

Figure 3. scBC performs best on a bicluster level

(A) Data simulation process. The parameter μ is computed by the multiplicative model $\mu = \mathbf{WZ}$. The prior edge information is generated along with $\mathbf{W}$. When generating $\mathbf{X}$, each element is generated from $\mathrm{NB}\left(r_j, \frac{1}{1+e^{-\mu_{ij}}}\right)$. To simulate different batches, we divided the dataset into 3 parts, with different intensities of noise. The implementation of dropout is to perform Bernoulli censoring. See STAR Methods for details.

(B–G) Performance of different methods under different conditions. For each plot, the x axis represents the dropout rate and the y axis represents the quantified performance (1-CE or F score). We ran 100 independent simulations for each setting; the data points represent mean value and the error bars represent the SD calculated across repeated simulations. (B) 1-CE of different methods under various dropout setting with simulated data scale: p=1000, n=300 and L=3. (C) F score of different methods under various dropout setting with simulated data scale: p=1000, n=300 and L=3. (D) 1-CE of different methods under various dropout setting with simulated data scale: p=3000, n=600 and L=4. (E) F score of different methods under various dropout setting with simulated data scale: p=3000, n=600 and L=4. (F) 1-CE of different methods under various dropout setting with simulated data scale: p=6000, n=1500 and L=5. (G) F score of different methods under various dropout setting with simulated data scale: p=6000, n=1500 and L=5.

Please cite this article in press as: Gong et al., Single-cell biclustering for cell-specific transcriptomic perturbation detection in AD progression, Cell Reports Methods (2024), https://doi.org/10.1016/j.crmeth.2024.100742



(legend on next page)

Please cite this article in press as: Gong et al., Single-cell biclustering for cell-specific transcriptomic perturbation detection in AD progression, Cell Reports Methods (2024), https://doi.org/10.1016/j.crmeth.2024.100742

demonstrated significant changes in Alzheimer's disease (AD) progression from early to late pathology. VSIG4 encodes a protein that is known to act as a negative regulator of T cell responses and is closely associated with impaired immune response. (Figure 5H). We also found that the cellular cation homeostasis pathway and synapse function are altered in the progression from normal to early pathology (Figure 5H).

Similar to excitatory neurons, pathways associated with kinase activity are also continuously altered throughout AD progression in microglia (Figure 5I). The cytokine-mediated signaling pathway is altered in early pathology (Figure 5I), which may be related to changes in the immune response in AD progression and is also in accordance with previous research.[32] Pathways related to gliogenesis and myelination also altered throughout the disease progression in microglia (Figure 5I), which is similar to astrocytes and excitatory neurons. Cell migration-related pathways are dysregulated in the late AD microglia (Figure 5I), which is also consistent with several studies[3,5,36,37] and largely related to microglial plaque clustering phenotypes, a phenomenon of inappropriate interactions with amyloid. The response to fatty acid becomes odd in early AD microglia (Figure 5I), which is also an indicator of lipid metabolism dysfunction. We also find cell chemotaxis becomes abnormal in late AD microglia (Figure 5I), indicating an inflammatory state in AD microglia.

We observed that in oligodendrocytes, the main changes in FGMs during AD progression occurred in myelination-related and synaptic signaling-related pathways (Figure 5J). Since memory preservation is thought to require new myelin formation, the impaired capacity of oligodendrocytes to adaptively monitor neural activity and facilitate myelin remodeling may govern cognitive decline in AD.[38] Moreover, synaptic signaling and axon development are critical for the transmission of excitement in the nervous system, and dysregulation of these processes can result in slower propagation of neural excitation. The changes in FGMs in oligodendrocytes are directly related to the reduction of nervous system excitability. Previous research also suggests that changes in oligodendrocytes may affect the function of other cells in the CNS.[39–42] Thus, targeting oligodendrocytes may be a promising strategy for the treatment of AD and other neurological disorders.

OPCs, which are distributed throughout gray and white matter, are thought to dynamically sense and modulate neural activity,[41] as oligodendrocytes do. Not surprisingly, then, pathways related to myelination and the ensheathment of neurons become abnormal along with the progression of AD (Figure 5K). Pathways related to ion transportation are also dysregulated, providing support for previous findings that genes related to ion channels are dysregulated in AD OPCs.[3,5] In addition, pathways related to kinase activity and cellular cation homeostasis are altered at the early stage in AD progression (Figure 5K).

Except for inhibitory neurons, which did not change significantly throughout disease progression, several other cell types exhibited specific FGM perturbations, highlighting the importance of single-cell analysis. Notably, we observed that pathways related to myelination and gliogenesis were more or less altered across all of these cell types, indicating similar alterations among AD-associated cells and suggesting that AD progression is largely related to dysregulation of this pathway, which was further confirmed in a recent study.[43]

Based on our findings, GBC exhibits relatively stable performance in FGM identification and ranks second only to scBC in disease-related datasets (Figure 2B). To further explore this, we conducted an analysis on the AD dataset using GBC with the same pipeline (STAR Methods). The results revealed that FGM perturbations were predominantly observed in the late-pathology stage across all cell types, except for OPCs (Figures S7–S12). This finding contradicts previous reports indicating widespread transcriptional changes occurring in the early stages of AD.[3] Furthermore, the enrichment analysis of perturbed FGMs identified by GBC yielded distinct results compared to scBC (Figures S7–S12) and was supported by little evidence from relevant studies. These results further highlight the unique advantages of scBC.

Sex-specific differential response in late AD microglia revealed by scBC

We observed that when the data were divided into pathology groups, cells from different sexes exhibited good merging in the no-pathology and early-pathology groups, but not in the late-pathology group (Figure 6A). As a result, we further stratified

Figure 4. Overview of the subsampled AD dataset

(A) Clinicopathological variables (columns) of 48 individuals (rows). Since the lower the value, the more serious the disease, here, we use its opposite number to be consistent with other indicators, so as to more intuitively show the differences between different pathology groups. Amyloid, overall amyloid level; cogn_global_lv, global cognitive function (last valid score); gpath, global AD pathology burden; gpath_3neocort, global measure of neocortical pathology; NFT, neurofibrillary tangle burden; plaq_n, neuritic plaque burden; Tangles, neuronal neurofibrillary tangle density.

(B) Uniform manifold approximation and projection (UMAP) visualization of all cells (n = 7,063) indicates cells from different donors of different pathological states are well blended. Color bar at the right represents the fraction composition of cells under different pathology.

(C) Same UMAP visualization as (B), but colored by sex. Color bar at right represents the fraction composition of cells of different sexes.

(D) The proportion of cells provided across individuals (columns). Bars represent the fraction of cells corresponding to each individual. Bar color indicates whether the corresponding value exceeds (blue-green) or does not exceed (rose red) the average value measured across all of the donors in the row. Red dashed line indicates the average.

(E) Fraction of cells of each type isolated from each individual (columns; n = 48).

(F) Fraction of cells of each type isolated across all (n = 48), no-pathology (n = 24), early-pathology (n = 15) and late-pathology (n = 9) individuals.

(G) Merge result between different biclusters. Gene sets from different biclusters in different pathology groups labeled with the same color are combined as a new FGM.

(H) The overlap of FGMs. The FGM marked with a solid black dot below the bar graph indicates that it is included in the comparison, and the FGM marked with a black transparent dot indicates that it is not included. For example, the first bar chart indicates that the number of genes appearing in FGM1 but not appearing in any other FGMs is 93. This result shows that the overlap between different FGMs is considerable.

CellPress
OPEN ACCESS

Cell Reports Methods
Article

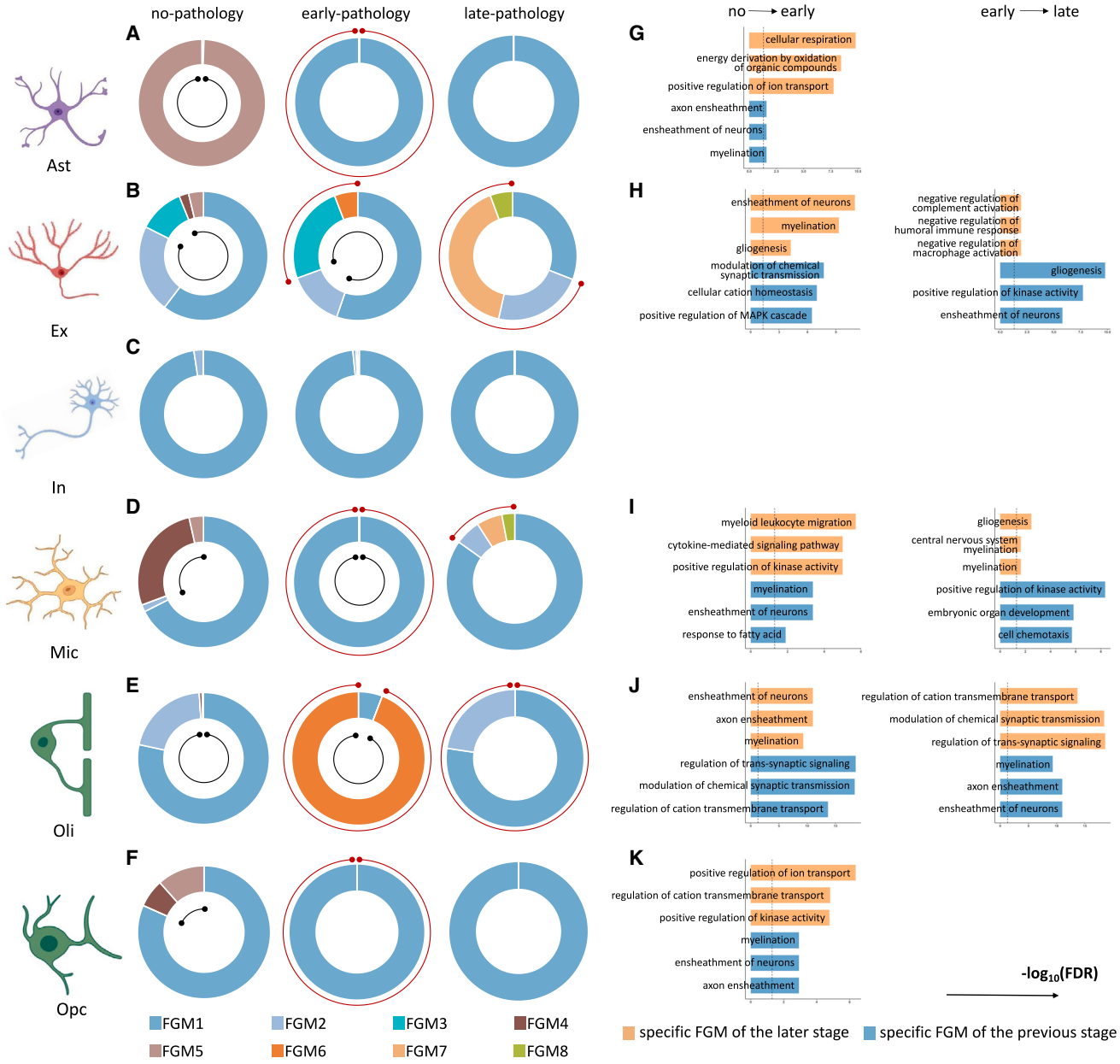


Figure 5. scBC uncover the pathway perturbation in AD progression

(A–F) Perturbation of FGM composition in each cell type during AD progression. Each pie chart quantifies the FGM composition of a cell under a specific progression condition. The outer red circles indicate FGMs whose composition is increasing compared to the previous stage. The inner black circle represents FGMs whose composition is decreasing compared to the later stage. (A) Perturbation of FGM composition in astrocytes. (B) Perturbation of FGM composition in excitatory neurons. (C) Perturbation of FGM composition in inhibitory neurons. (D) Perturbation of FGM composition in microglia. (E) Perturbation of FGM composition in oligodendrocytes. (F) Perturbation of FGM composition in oligodendrocyte precursor cells.

(G–K) Results of enrichment analysis of FGMs altered in 2 phases. Orange represents the specific FGM of the later stage, and blue represents the specific FGM of the previous stage, both representing the set of genes that are perturbated during the progression. (G) Results of enrichment analysis of FGMs altered in astrocytes. (H) Results of enrichment analysis of FGMs altered in excitatory neurons. (I) Results of enrichment analysis of FGMs altered in microglia. (J) Results of enrichment analysis of FGMs altered in oligodendrocytes. (K) Results of enrichment analysis of FGMs altered in oligodendrocyte precursor cells.

the cells by sex to examine the differences in FGM composition between sexes using the scBC results. The stratified results revealed that in the no-pathology and early-pathology groups, FGM compositions remained consistently similar across sexes and aligned with the findings from the combined analysis (Figure S13). However, in the late-pathology group, most FGM compositions were consistent between different sex groups, except for microglia (Figure 6B). Although there appeared to

Please cite this article in press as: Gong et al., Single-cell biclustering for cell-specific transcriptomic perturbation detection in AD progression, Cell Reports Methods (2024), https://doi.org/10.1016/j.crmeth.2024.100742

CellPress
OPEN ACCESS

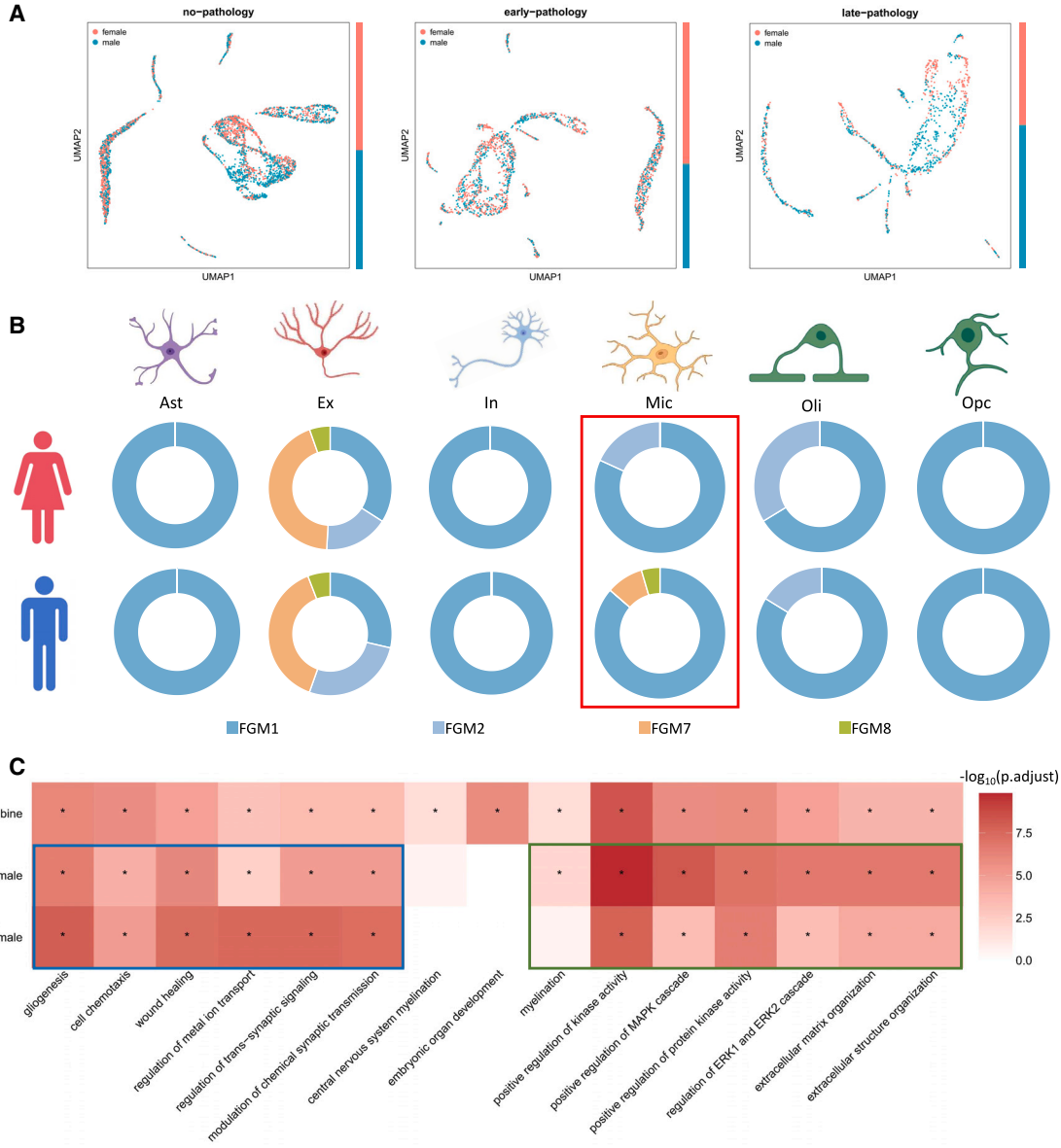


Figure 6. Sex-specific differential response in late AD microglia revealed by scBC

(A) UMAP visualization of cells from different pathology groups and colored by sex. Color bar at right represents the fraction composition of cells from different sexes.

(B) FGM composition in each cell type in late-pathology group, stratified by sex. Each pie chart quantifies the FGM composition of a cell type. Microglia represented an obvious difference of FGM composition in different sexes and is highlighted in the plot.

(C) Heatmap of −log10 transformed p values (BH adjusted) for the top enriched pathways from sex-specific perturbed FGMs along with the previous combined result. The blue box indicates perturbed pathways that are more significant in the male group, whereas the green box indicates perturbed pathways that are more significant in the female group. Asterisk in the tile denotes significance (adjusted $p < 0.05$).

be variations in the proportion of FGMs in oligodendrocytes, the perturbed FGMs compared to the previous pathology stage remained the same, thus not influencing the enrichment results. The sex-specific differential response in AD microglia is also widely reported in previous research,[44–46] proving the reliability of the scBC findings. To further investigate sex-specific pathway perturbations in late AD microglia, we performed an enrichment analysis on the perturbed FGMs in different sexes and selected the top enriched pathways in conjunction with the previously combined ones for comparison.

It is intriguing to observe that the perturbed pathway of embryonic organ development, which initially seemed unrelated to the function of microglia, was not detected in either sex group in the combined analysis (Figure 6C). In addition, the differential enrichment of pathways between the different sex groups emphasizes the importance of conducting a stratified

Please cite this article in press as: Gong et al., Single-cell biclustering for cell-specific transcriptomic perturbation detection in AD progression, Cell Reports Methods (2024), https://doi.org/10.1016/j.crmeth.2024.100742

analysis based on sex (Figure 6C). Notably, a distinct response in late-stage AD microglia is observed, with signaling-related pathways (e.g., cell chemotaxis, regulation of metal ion transport, regulation of *trans*-synaptic signaling, and modulation of chemical synaptic transmission) primarily being perturbed in male individuals, whereas kinase activity-related pathways (including positive regulation of kinase activity, positive regulation of the MAPK cascade, positive regulation of protein kinase activity, and regulation of the ERK1 and ERK2 cascade) are mainly perturbed in female individuals (Figure 6C). We believe that these results provided by scBC offer valuable guidance for future investigations to unravel the underlying mechanisms driving the sexual dimorphism observed in AD pathology.

DISCUSSION

Molecular biomarkers have been widely used in clinical practice to identify diseases, but they often suffer from low coverage and high false positive or false negative rates, limiting their further application.[47] Network biomarkers, also known as module biomarkers, have attracted attention as a more robust form of biomarker than individual molecules for characterizing diseases.[48,49] This is particularly important for analyzing single-cell data, which are inherently more complex than bulk tissue data due to the heterogeneity of individual cells within a sample. However, network biomarkers are usually cell specific and may change during disease progression. To detect cell-specific network biomarkers, we developed scBC, a single-cell Bayesian biclustering method that combines VAE for batch removal and data imputation with matrix factorization-based Bayesian biclustering for using known biological information. Our method outperforms other state-of-the-art methods in finding FGMs, discovering functionally related cell groups, and detecting cell-gene correlation patterns in highly heterogeneous scRNA-seq datasets and simulated data. This makes scBC well suited for analyzing diseases with multifactorial etiologies whose functionally related potential cell groups are finely divided, cell-type conditioned gene co-expression patterns are complicated, and cell-gene correlation changes throughout the disease progression.

In this study, we applied scBC to an snRNA AD dataset to explore how the transcriptional functional modules of each cell type change as the disease progresses. Our results further confirmed the complex interplay of virtually every major brain cell type in AD.[3,34] We found that FGM composition largely changed in astrocytes and oligodendrocyte precursor cells before individuals developed severe pathological features. However, inhibitory neurons showed minimal changes in FGM throughout disease progression, indicating that this cell type does not have many alterations in transcriptional patterns during AD progression. A consistent FGM perturbation across all other cell types, except inhibitory neurons, was the alteration in pathways related to myelination and gliogenesis, suggesting that this pathway may play a decisive role in the progression of AD.

Specific to each cell type, energy metabolism and ion transporters are dysregulated in AD astrocytes, indicating the inflammatory state of the brain following injury and neurodegeneration. The perturbation of FGM composition in excitatory neurons from normal to early pathology is very subtle compared to the change from the early stage to the late stage, indicating that the rate at which the cells become abnormal may be slow at first and then fast. Another continuous change in excitatory neurons in AD progression is the general dysregulation in kinase activity, which is closely related to neuronal DNA damage. In addition, immune response, cellular cation homeostasis, and synapse function are altered in AD excitatory neurons. Microglia shares a similar alteration in kinase activity with excitatory neurons throughout AD progression. Pathways such as immune response-related cytokine-mediated signaling, amyloid interaction-related cell migration, and lipid metabolism-related fatty acid response are dysregulated in the early AD microglia. Cell chemotaxis becomes abnormal in late AD microglia, indicating an inflammatory state in AD microglia. The oligodendrocyte is a cell that needs to be focused on more for disease treatment since FGM perturbations in such a cell type are mainly concentrated in myelination-related and synaptic signaling-related pathways, directly related to the reduction of the excitability of the nervous system. Pathways related to ion transportation, kinase activity, and cellular cation homeostasis are dysregulated in oligodendrocyte precursor cells. Finally, sex-specific differential response in AD microglia is also reported. Specifically, signaling-related pathways are primarily perturbed in male individuals, whereas kinase activity-related pathways are mainly perturbed in female individuals.

Limitations of the study

In the context of high-throughput sequencing data, network biomarker-based analytical methods preserve the complex co-expression or co-regulation patterns in the gene module and are more robust to the analysis of complex diseases. We believe that scBC, as a technique for cell-specific network biomarker detection, creates an opportunity for effectively delineating mechanisms of complex diseases at single-cell resolution, providing advice on the treatment of such diseases. However, although the network biomarker may contain complex coexpression or co-regulation patterns, its internal precise and quantitative regulatory relationship has not been clarified. Future research can focus on the explanation of the regulatory relationship within the cell-specific network structure, so as to have a more accurate inference on the principle of FGM perturbations during disease progression. In addition, although the application of our method to AD scRNA data analysis yielded consistent conclusions with previous studies, it is important to note that most of these supporting conclusions were derived from computational analysis rather than experimental validation. Furthermore, given the presentation of numerous hyperparameters in scBC, although we have set them to reasonable defaults, exploring alternative hyperparameter tuning approaches may lead to improved model fit and more appropriate inference in certain cases. Lastly, matrix factorization-based Bayesian optimization procedure is also time-consuming, especially when the dataset is extremely large. An unbiased subsampling procedure is crucial when dealing with extremely large datasets. How to speed up calculations while ensuring algorithm accuracy is also of interest for future research.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Details for data reconstruction using variational inference
  - Model training at learning stage
  - Bayesian biclustering incorporate biological information
  - MAP estimation for biclustering result
  - Strong classification of cell group for different biclustering methods
  - Datasets and preprocessing
  - Matching of biclusters when analyzing AD dataset
  - FGM perturbation during AD progression and enrichment analysis
  - Analyzing AD dataset using GBC
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Comparison of FGM detection
  - Criterions for clustering performance
  - Metrices for biclustering comparison

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2024.100742.

ACKNOWLEDGMENTS

The research is supported partly by the National Natural Science Foundation of China (11901387 to Y.Z. and 12171318 to Z.Y.) and Shanghai Jiao Tong University "Jiaotong Star" Plan Medical Engineering Cross Research Project (20230103 to Y.Z.). The computations in this paper were run on the Siyuan-1 cluster supported by the Center for High Performance Computing, Shanghai Jiao Tong University.

AUTHOR CONTRIBUTIONS

Y.G. performed the research, analyzed the data, and wrote the original manuscript. J.X. participated in the data collection. M.W. provided practical suggestions and technical instructions for the research. Z.Y. and Y.Z. supervised the research. All of the authors discussed and revised the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 23, 2023
Revised: October 30, 2023
Accepted: March 7, 2024
Published: March 29, 2024

REFERENCES

1. Shi, F., and Huang, H. (2017). Identifying Cell Subpopulations and Their Genetic Drivers from Single-Cell RNA-Seq Data Using a Biclustering Approach. J. Comput. Biol. *24*, 663–674. https://doi.org/10.1089/cmb.2017.0049.
2. Eisenberg, E., and Levanon, E.Y. (2013). Human housekeeping genes, revisited. Trends Genet. *29*, 569–574. https://doi.org/10.1016/j.tig.2013.05.010.
3. Mathys, H., Davila-Velderrain, J., Peng, Z., Gao, F., Mohammadi, S., Young, J.Z., Menon, M., He, L., Abdurrob, F., Jiang, X., et al. (2019). Single-cell transcriptomic analysis of Alzheimer's disease. Nature *570*, 332–337. https://doi.org/10.1038/s41586-019-1195-2.
4. Lake, B.B., Chen, S., Sos, B.C., Fan, J., Kaeser, G.E., Yung, Y.C., Duong, T.E., Gao, D., Chun, J., Kharchenko, P.V., and Zhang, K. (2018). Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. Nat. Biotechnol. *36*, 70–80. https://doi.org/10.1038/nbt.4038.
5. Grubman, A., Chew, G., Ouyang, J.F., Sun, G., Choo, X.Y., McLean, C., Simmons, R.K., Buckberry, S., Vargas-Landin, D.B., Poppe, D., et al. (2019). A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. Nat. Neurosci. *22*, 2087–2097. https://doi.org/10.1038/s41593-019-0539-4.
6. Habib, N., McCabe, C., Medina, S., Varshavsky, M., Kitsberg, D., Dvir-Szternfeld, R., Green, G., Dionne, D., Nguyen, L., Marshall, J.L., et al. (2020). Disease-associated astrocytes in Alzheimer's disease and aging. Nat. Neurosci. *23*, 701–706. https://doi.org/10.1038/s41593-020-0624-8.
7. Roussarie, J.P., Yao, V., Rodriguez-Rodriguez, P., Oughtred, R., Rust, J., Plautz, Z., Kasturia, S., Albornoz, C., Wang, W., Schmidt, E.F., et al. (2020). Selective Neuronal Vulnerability in Alzheimer's Disease: A Network-Based Analysis. Neuron *107*, 821–835.e12. https://doi.org/10.1016/j.neuron.2020.06.010.
8. Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinf. *9*, 559. https://doi.org/10.1186/1471-2105-9-559.
9. Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. Stat. Appl. Genet. Mol. Biol. *4*, Article17. https://doi.org/10.2202/1544-6115.1128.
10. Ghazalpour, A., Bennett, B., Petyuk, V.A., Orozco, L., Hagopian, R., Mungrue, I.N., Farber, C.R., Sinsheimer, J., Kang, H.M., Furlotte, N., et al. (2011). Comparative analysis of proteome and transcriptome variation in mouse. PLoS Genet. *7*, e1001393. https://doi.org/10.1371/journal.pgen.1001393.
11. Zhang, B., Gaiteri, C., Bodea, L.G., Wang, Z., McElwee, J., Podtelezhnikov, A.A., Zhang, C., Xie, T., Tran, L., Dobrin, R., et al. (2013). Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. Cell *153*, 707–720. https://doi.org/10.1016/j.cell.2013.03.030.
12. Risso, D., Ngai, J., Speed, T.P., and Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. Nat. Biotechnol. *32*, 896–902. https://doi.org/10.1038/nbt.2931.
13. Hochreiter, S., Bodenhofer, U., Heusel, M., Mayr, A., Mitterecker, A., Kasim, A., Khamiakova, T., Van Sanden, S., Lin, D., Talloen, W., et al. (2010). FABIA: factor analysis for bicluster acquisition. Bioinformatics *26*, 1520–1527. https://doi.org/10.1093/bioinformatics/btq227.
14. Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., Geman, D., Baggerly, K., and Irizarry, R.A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. Nat. Rev. Genet. *11*, 733–739. https://doi.org/10.1038/nrg2825.
15. Hu, Z., Zu, S., and Liu, J.S. (2020). SIMPLEs: a single-cell RNA sequencing imputation strategy preserving gene modules and cell clusters variation. NAR Genom. Bioinform. *2*, lqaa077. https://doi.org/10.1093/nargab/lqaa077.
16. Fang, Q., Su, D., Ng, W., and Feng, J. (2021). An Effective Biclustering-Based Framework for Identifying Cell Subpopulations From scRNA-seq Data. IEEE/ACM Trans. Comput. Biol. Bioinform. *18*, 2249–2260. https://doi.org/10.1109/TCBB.2020.2979717.

17. Xie, J., Ma, A., Zhang, Y., Liu, B., Cao, S., Wang, C., Xu, J., Zhang, C., and Ma, Q. (2020). QUBIC2: a novel and robust biclustering algorithm for analyses and interpretation of large-scale RNA-Seq data. Bioinformatics *36*, 1143–1149. https://doi.org/10.1093/bioinformatics/btz692.

18. Zhong, Y., and Huang, J.Z. (2022). Biclustering via structured regularized matrix decomposition. Stat. Comput. *32*, 37. https://doi.org/10.1007/s11222-022-10095-1.

19. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nat. Methods *15*, 1053–1058. https://doi.org/10.1038/s41592-018-0229-2.

20. Li, Z., Chang, C., Kundu, S., and Long, Q. (2020). Bayesian generalized biclustering analysis via adaptive structured shrinkage. Biostatistics *21*, 610–624. https://doi.org/10.1093/biostatistics/kxy081.

21. Cheng, Y., and Church, G.M. (2000). Biclustering of expression data. Proc. Int. Conf. Intell. Syst. Mol. Biol. *8*, 93–103.

22. Murali, T.M., and Kasif, S. (2003). Extracting conserved gene expression motifs from gene expression data Pacific Symposium on Biocomputing. In Pacific Symposium on Biocomputing, pp. 77–88.

23. Prelić, A., Bleuler, S., Zimmermann, P., Wille, A., Bühlmann, P., Gruissem, W., Hennig, L., Thiele, L., and Zitzler, E. (2006). A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics *22*, 1122–1129. https://doi.org/10.1093/bioinformatics/btl060.

24. Caldas, J., and Kaski, S. (2008). Bayesian Biclustering with the Plaid Model (Machine Learn Sign P), pp. 291–296. https://doi.org/10.1109/MlSp.2008.4685495.

25. Xu, J., Xu, J., Meng, Y., Lu, C., Cai, L., Zeng, X., Nussinov, R., and Cheng, F. (2023). Graph embedding and Gaussian mixture variational autoencoder network for end-to-end analysis of single-cell RNA sequencing data. Cell Rep. Methods *3*, 100382. https://doi.org/10.1016/j.crmeth.2022.100382.

26. Dongen, S.V. (2000). Graph clustering by flow simulation. PhD Thesis (University of Utrecht).

27. Milligan, G.W., and Cooper, M.C. (1986). A STUDY OF THE COMPARABILITY OF EXTERNAL CRITERIA FOR HIERARCHICAL CLUSTER-ANALYSIS. Multivariate Behav. Res. *21*, 441–458. https://doi.org/10.1207/s15327906mbr2104_5.

28. Santos, J.M., and Embrechts, M. (2009). On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification. Held in Limassol (CYPRUS), pp. 175–+. 2009 Sep 14-17.

29. Fowlkes, E.B., and Mallows, C.L. (1983). A method for comparing two hierarchical clusterings. J. Am. Stat. Assoc. *78*, 553–569.

30. Strehl, A., and Ghosh, J. (2003). Cluster ensembles- a knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res. *3*, 583–617. https://doi.org/10.1162/153244303321897735.

31. Rahaman, M.A., Rodrigue, A., Glahn, D., Turner, J., and Calhoun, V. (2021). Shared sets of correlated polygenic risk scores and voxel-wise grey matter across multiple traits identified via bi-clustering. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. *2021*, 2201–2206. https://doi.org/10.1109/EMBC46164.2021.9630825.

32. Murdock, M.H., and Tsai, L.H. (2023). Insights into Alzheimer's disease from single-cell genomic approaches. Nat. Neurosci. *26*, 181–195. https://doi.org/10.1038/s41593-022-01222-2.

33. Hasel, P., Rose, I.V.L., Sadick, J.S., Kim, R.D., and Liddelow, S.A. (2021). Neuroinflammatory astrocyte subtypes in the mouse brain. Nat. Neurosci. *24*, 1475–1487. https://doi.org/10.1038/s41593-021-00905-6.

34. Blanchard, J.W., Akay, L.A., Davila-Velderrain, J., von Maydell, D., Mathys, H., Davidson, S.M., Effenberger, A., Chen, C.Y., Maner-Smith, K., Hajjar, I., et al. (2022). APOE4 impairs myelination via cholesterol dysregulation in oligodendrocytes. Nature *611*, 769–779. https://doi.org/10.1038/s41586-022-05439-w.

35. Welch, G., and Tsai, L.H. (2022). Mechanisms of DNA damage-mediated neurotoxicity in neurodegenerative disease. EMBO Rep. *23*, e54217. https://doi.org/10.15252/embr.202154217.

36. Zhou, Y., Song, W.M., Andhey, P.S., Swain, A., Levy, T., Miller, K.R., Poliani, P.L., Cominelli, M., Grover, S., Gilfillan, S., et al. (2020). Author Correction: Human and mouse single-nucleus transcriptomics reveal TREM2-dependent and TREM2-independent cellular responses in Alzheimer's disease. Nat. Med. *26*, 981. https://doi.org/10.1038/s41591-020-0922-4.

37. Lau, S.F., Cao, H., Fu, A.K.Y., and Ip, N.Y. (2020). Single-nucleus transcriptome analysis reveals dysregulation of angiogenic endothelial cells and neuroprotective glia in Alzheimer's disease. Proc. Natl. Acad. Sci. USA *117*, 25800–25809. https://doi.org/10.1073/pnas.2008762117.

38. Pan, S., Mayoral, S.R., Choi, H.S., Chan, J.R., and Kheirbek, M.A. (2020). Preservation of a remote fear memory requires new myelin formation. Nat. Neurosci. *23*, 487–499. https://doi.org/10.1038/s41593-019-0582-1.

39. Fancy, S.P.J., Chan, J.R., Baranzini, S.E., Franklin, R.J.M., and Rowitch, D.H. (2011). Myelin regeneration: a recapitulation of development? Annu. Rev. Neurosci. *34*, 21–43. https://doi.org/10.1146/annurev-neuro-061010-113629.

40. Franklin, R.J.M., and Goldman, S.A. (2015). Glia Disease and Repair-Remyelination. Cold Spring Harb. Perspect. Biol. *7*, a020594. https://doi.org/10.1101/cshperspect.a020594.

41. Káradóttir, R., Hamilton, N.B., Bakiri, Y., and Attwell, D. (2008). Spiking and nonspiking classes of oligodendrocyte precursor glia in CNS white matter. Nat. Neurosci. *11*, 450–456. https://doi.org/10.1038/nn2060.

42. Mitew, S., Hay, C.M., Peckham, H., Xiao, J., Koenning, M., and Emery, B. (2014). Mechanisms regulating the development of oligodendrocytes and central nervous system myelin. Neuroscience *276*, 29–47. https://doi.org/10.1016/j.neuroscience.2013.11.029.

43. Depp, C., Sun, T., Sasmita, A.O., Spieth, L., Berghoff, S.A., Nazarenko, T., Overhoff, K., Steixner-Kumar, A.A., Subramanian, S., Arinrad, S., et al. (2023). Myelin dysfunction drives amyloid-beta deposition in models of Alzheimer's disease. Nature *618*, 349–357. https://doi.org/10.1038/s41586-023-06120-6.

44. Mhatre, S.D., Tsai, C.A., Rubin, A.J., James, M.L., and Andreasson, K.I. (2015). Microglial malfunction: the third rail in the development of Alzheimer's disease. Trends Neurosci. *38*, 621–636. https://doi.org/10.1016/j.tins.2015.08.006.

45. Villa, A., Gelosa, P., Castiglioni, L., Cimino, M., Rizzi, N., Pepe, G., Lolli, F., Marcello, E., Sironi, L., Vegeto, E., and Maggi, A. (2018). Sex-Specific Features of Microglia from Adult Mice. Cell Rep. *23*, 3501–3511. https://doi.org/10.1016/j.celrep.2018.05.048.

46. Mosher, K.I., and Wyss-Coray, T. (2014). Microglial dysfunction in brain aging and Alzheimer's disease. Biochem. Pharmacol. *88*, 594–604. https://doi.org/10.1016/j.bcp.2014.01.008.

47. Liu, R., Wang, X., Aihara, K., and Chen, L. (2014). Early diagnosis of complex diseases by molecular biomarkers, network biomarkers, and dynamical network biomarkers. Med. Res. Rev. *34*, 455–478. https://doi.org/10.1002/med.21293.

48. Jin, G., Zhou, X., Wang, H., Zhao, H., Cui, K., Zhang, X.S., Chen, L., Hazen, S.L., Li, K., and Wong, S.T.C. (2008). The knowledge-integrated network biomarkers discovery for Major Adverse Cardiac Events. J. Proteome Res. *7*, 4013–4021. https://doi.org/10.1021/pr8002886.

49. Ideker, T., and Sharan, R. (2008). Protein networks in disease. Genome Res. *18*, 644–652. https://doi.org/10.1101/gr.071852.107.

50. Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049. https://doi.org/10.1038/ncomms14049.

51. Kim, N., Kim, H.K., Lee, K., Hong, Y., Cho, J.H., Choi, J.W., Lee, J.I., Suh, Y.L., Ku, B.M., Eum, H.H., et al. (2020). Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. Nat. Commun. *11*, 2285. https://doi.org/10.1038/s41467-020-16164-1.

52. Chung, W., Eum, H.H., Lee, H.O., Lee, K.M., Lee, H.B., Kim, K.T., Ryu, H.S., Kim, S., Lee, J.E., Park, Y.H., et al. (2017). Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. Nat. Commun. *8*, 15081. https://doi.org/10.1038/ncomms15081.

53. Li, Z., Safo, S.E., and Long, Q. (2017). Incorporating biological information in sparse principal component analysis with application to genomic data. BMC Bioinf. *18*, 332. https://doi.org/10.1186/s12859-017-1740-7.

54. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J.P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. Nat. Commun. *9*, 284. https://doi.org/10.1038/s41467-017-02554-5.

55. Grün, D., Kester, L., and van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. Nat. Methods *11*, 637–640. https://doi.org/10.1038/Nmeth.2930.

56. Svensson, V., Natarajan, K.N., Ly, L.H., Miragaia, R.J., Labalette, C., Macaulay, I.C., Cvejic, A., and Teichmann, S.A. (2017). Power analysis of single-cell RNA-sequencing experiments. Nat. Methods *14*, 381–387. https://doi.org/10.1038/nmeth.4220.

57. Lun, A.T.L., McCarthy, D.J., and Marioni, J.C. (2016). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Res. *5*, 2122. https://doi.org/10.12688/f1000research.9501.2.

58. Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A.K., Slichter, C.K., Miller, H.W., McElrath, M.J., Prlic, M., et al. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. *16*, 278. https://doi.org/10.1186/s13059-015-0844-5.

59. Kingma, D.P., and Welling, M. (2013). Auto-Encoding Variational Bayes. Preprint at arXiv. https://doi.org/10.48550/arXiv.1312.6114.

60. Sonderby, C.K., Raiko, T., Maaloe, L., Sonderby, S.K., and Winther, O. (2016). Ladder Variational Autoencoders. Adv Neur *29*.

61. Li, C., and Li, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. Bioinformatics *24*, 1175–1182. https://doi.org/10.1093/bioinformatics/btn081.

62. Zhao, Y., Chung, M., Johnson, B.A., Moreno, C.S., and Long, Q. (2016). Hierarchical Feature Selection Incorporating Known and Novel Biological Information: Identifying Genomic Features Related to Prostate Cancer Recurrence. J. Am. Stat. Assoc. *111*, 1427–1439. https://doi.org/10.1080/01621459.2016.1164051.

63. Safo, S.E., Li, S., and Long, Q. (2018). Integrative analysis of transcriptomic and metabolomic data via sparse canonical correlation analysis with incorporation of biological information. Biometrics *74*, 300–312. https://doi.org/10.1111/biom.12715.

64. Chang, C., Kundu, S., and Long, Q. (2018). Scalable Bayesian variable selection for structured high-dimensional data. Biometrics *74*, 1372–1382. https://doi.org/10.1111/biom.12882.

65. Polson, N.G., Scott, J.G., and Windle, J. (2013). Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables. J. Am. Stat. Assoc. *108*, 1339–1349. https://doi.org/10.1080/01621459.2013.829001.

66. Chang, C., and Tsay, R.S. (2010). Estimation of covariance matrix via the sparse Cholesky factor with lasso. J. Stat. Plann. Inference *140*, 3858–3873. https://doi.org/10.1016/j.jspi.2010.04.048.

67. Patrikainen, A., and Meila, M. (2006). Comparing subspace clusterings. Ieee T Knowl Data En *18*, 902–916. https://doi.org/10.1109/Tkde.2006.106.

Please cite this article in press as: Gong et al., Single-cell biclustering for cell-specific transcriptomic perturbation detection in AD progression, Cell Reports Methods (2024), https://doi.org/10.1016/j.crmeth.2024.100742

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
HEART	Heart Cell Atlas	https://github.com/YosefLab/scVI-data/blob/master/hca_subsampled_20k.h5ad
PBMC	Zheng, G.X. et al.[50]	https://github.com/YosefLab/scVI-data/raw/master/PurifiedPBMCDataset.h5ad
Lung adenocarcinoma (LUAD)	Kim, N. et al.[51]	GEO: GSE131907
breast cancer (BC)	Chung, W. et al.[52]	GEO: GSE75688
Alzheimer's disease (AD)	ROSMAP	Synapse: syn18485175
Software and algorithms		
CC	Cheng, Y. and G.M. Church[21]	https://github.com/cran/biclust
xMotifs	Murali, T.M. and S. Kasif[22]	https://github.com/cran/biclust
FABIA	Hochreiter, S. et al.[13]	https://new.bioconductor.org/packages/release/bioc/html/fabia.html
Bimax	Prelic, A. et al.	https://github.com/cran/biclust
plaid	Caldas, J. and S. Kaski[24]	https://github.com/cran/biclust
GBC	Li, Z. et al.[20,53]	https://github.com/ziyili20/GBC
QUBIC2	Xie, J. et al.[17]	https://github.com/maqin2001/qubic2
DivBiclust	Fang, Q. et al.[16]	https://github.com/Qiong-Fang/DivBiclust
autoCell	Xu, J. et al.[25]	https://github.com/ChengF-Lab/autoCell
scBC	This paper	https://github.com/GYQ-form/scBC https://doi.org/10.5281/zenodo.10777594

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Yue Zhang (yue.zhang@sjtu.edu.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyzes existing, publicly available data. The accession numbers for the datasets are listed in the key resources table.
- Our scBC method is available as a Python package on PyPI at https://pypi.org/project/scBC, free for academic use. All original code has been deposited at https://github.com/GYQ-form/scBC and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon request.

METHOD DETAILS

Details for data reconstruction using variational inference

Taking advantage of recent work by Romain et al.,[19] here we also adopt the idea of using variational inference to estimate the posterior distribution for the low-dimensional, latent variables z_n for each cell n which should reflect biological differences among cells. To remove the nuisance variation due to technique factors such as batch effects, it's reasonable to model the sampling distribution conditioned on the batch annotations s_n[54,55]. That is, the observed expression x_n^g of each gene g in each cell n is drawn from $p(x_n^g|z_n, s_n)$. There has been some discussion about how to model the scRNA-seq data. Zero-inflated

Please cite this article in press as: Gong et al., Single-cell biclustering for cell-specific transcriptomic perturbation detection in AD progression, Cell Reports Methods (2024), https://doi.org/10.1016/j.crmeth.2024.100742

CellPress
OPEN ACCESS

negative binomial distribution (ZINB) or negative binomial (NB) distribution are deemed as the better choice.[55–58] To model the data generation from a ZINB or NB distribution, we use a hierarchical probabilistic model for data generating process:

$$z_n \sim N(0, \boldsymbol{I})$$

$$\rho_n \sim f_{expect}(z_n, s_n)$$

$$w_n^g \sim Gamma\left(\rho_n^g, \theta^g\right)$$

$$y_n^g \sim Poisson\left(l_n w_n^g\right)$$

$$h_n^g \sim Bernoulli(f_{drop}(z_n, s_n))$$

The subscript n denotes the representation of n_{th} cell, which typically is a multi-dimensional vector. l_n is a parameter strongly correlated with and decide the library size of cell n. We use superscript annotation (for example, ρ_n^g) to refer to a single entry that corresponds to a specific gene g. The parameter $\theta \subset \mathbb{R}_+^G$ denotes a gene-specific inverse dispersion, which can be estimated via variational Bayesian inference. Here z_n is the low-dimensional, latent variable for cell n. We use a standard multivariate normal prior for z because it can be reparametrized in a differentiable way into any arbitrary multivariate Gaussian random variable, which is extremely helpful in the inference process. We denote B as number of batches, then f_{expect} is a neuron network which maps the latent space and batch annotations of each cell back to the full dimension of the gene expression: $\mathbb{R}^{d+1} \times \{0,1\}^B \rightarrow \mathbb{R}^G$. At the generating stage, f_{expect} is constrained by a softmax activation function at the last layer so that each element of ρ_n sum up to 1 during inference. Therefore, ρ_n denotes the mean proportion of transcripts expressed across all genes. $f_{drop}(z_n, s_n)$ is also a neuron network which maps the latent space and batch annotations of each cell to their respective dropout probabilities. w_n^g and y_n^g are two intermediate variable and it can be shown that through this process, h_n^g is an r.v. following ZINB distribution[55] with mean $l_n \rho_n^g$, gene-specific dispersion θ^g and zero-inflation probability $f_{drop}(z_n, s_n)$ (See proof below).

When we conduct data reconstruction to get the batch-removal, imputed gene expression data, we only take advantage of the intermediate variable ρ_n and scale it to our expected library size. That is, multiplying it by a given parameter, which we just use the empirical library size (total number of transcripts per cell) of each cell throughout our experiments. But one should notice we can re-scale it to any expected library size if additional information is given.

Marginal distribution of generation distribution

Through the generation procedure, we can model the scRNA-seq data either as ZINB or NB distribution. The proof is as following:

First, take r to be the gene-specific shape parameter of a Gamma variable w and $\frac{p}{1-p}$ to be its scale parameter, use a scalar $\lambda \in \mathbb{R}^+$, then the count variable $y|w \sim$ Poisson (λw) has a negative binomial marginal distribution with mean $r\lambda\frac{p}{1-p}$:

$$\begin{aligned} p(y) &= \int p(y|w)p(w)dw \\ &= \int \frac{e^{-\lambda w}\lambda^y w^y}{\Gamma(y+1)} \frac{w^{r-1}e^{-w\left(\frac{1}{p}-1\right)}}{p^r\Gamma(r)}(1-p)^r dw \\ &= \frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)}\left(\frac{1-p}{1-p+\lambda p}\right)^r\left(\frac{p\lambda}{1-p+\lambda p}\right)^y \end{aligned}$$

Second, multiplication by zero to y_n^g can be formally encoded as a mixture between a point-mass at zero and the original distribution of y_n^g. Consequently, the conditional $p(x_n^g|z_n, s_n)$ is a zero-inflated negative binomial with probability mass function (for simplicity, we ignore the subscript n):

$$p(x^g = 0|z, l, s) = f_{drop}(z,s)^g + \left(1 - f_{drop}(z,s)^g\right)\left(\frac{\theta^g}{\theta^g + l}\right)^{f_{expect}(z,s)}$$

$$p(x^g = y|z, l, s) = \left(1 - f_{drop}(z,s)^g\right)\frac{\Gamma(y + f_{expect}(z,s))}{\Gamma(y+1)\Gamma(f_{expect}(z,s))}\left(\frac{\theta^g}{\theta^g + l}\right)^{f_{expect}(z,s)}\left(\frac{l}{\theta^g + l}\right)^y,\ y \in N*$$

Model training at learning stage

In our VAE architecture, we denote variational parameters as φ and generative parameters as θ. Here we introduce a recognition model $q_\varphi(z_n|x_n, s_n)$: an approximation to the intractable true posterior $p_\theta(z_n|x_n, s_n)$. The marginal likelihood can be written as:

$$\log p_\theta(x_n|s_n) = D_{KL}(q_\varphi(z_n|x_n, s_n)||p_\theta(z_n|x_n, s_n)) + L(\theta, \varphi;\ x_n)$$

Where $L(\theta, \varphi; x_n) = E_{q_\varphi(z_n|x_n,s_n)}[-\log q_\varphi(z_n|x_n, s_n) + \log p_\theta(z_n, x_n|s_n)]$. Since the KL-divergence is always non-negative. We have:

$$\log p_\theta(x_n|s_n) \geq E_{q_\varphi(z_n|x_n,s_n)}[-\log q_\varphi(z_n|x_n, s_n) + \log p_\theta(z_n, x_n|s_n)]$$

The evidence lower bound (ELBO) $L(\theta, \varphi; x_n)$ can also be written as:

$$L(\theta, \varphi; x_n) = E_{q_\varphi(z_n|x_n,s_n)}[\log p_\theta(x_n|z_n, s_n)] - D_{KL}(q_\varphi(z_n|x_n, s_n)\|p_\theta(z_n|s_n))$$

Optimizing the ELBO means optimizing both the variational parameters φ and generative parameters θ at the same time. Assuming the true latent variable z_n is batch-free (independent with batch annotation s_n) and the prior follows standard multivariate Gaussian distribution, we can get the closed-form expression of the derivative of $D_{KL}(q_\varphi(z_n|x_n, s_n)\|p_\theta(z_n|s_n))$. To get the low-variance Monte Carlo estimation of the gradient of term $E_{q_\varphi(z_n|x_n,s_n)}[\log p_\theta(x_n|z_n, s_n)]$, we use the reparameterization trick in the learning stage[59]:

$$\tilde{E}_{q_\varphi(z_n|x_n,s_n)}[\log p_\theta(x_n|z_n, s_n)] \cong \frac{1}{L}\sum_{l=1}^{L} \log p_\theta\left(x_n|g_\varphi\left(\epsilon^l, x_n\right), s_n\right)$$

Where $g_\varphi(\epsilon^l, x_n)$ is a differentiable transformation to reparameterize the random variable $z_n \sim q_\varphi(z_n|x_n, s_n)$ and $\epsilon \sim p(\epsilon)$ is an auxiliary noise variable.

For a single data point x_n (cell n) we have:

$$\tilde{L}(\theta, \varphi; x_n) = \frac{1}{L}\sum_{l=1}^{L} \log p_\theta\left(x_n|z^{(l)}, s_n\right) - D_{KL}(q_\varphi(z_n|x_n, s_n)\|p_\theta(z_n))$$

Where $z^{(l)} = g_\varphi(\epsilon^{(l)}, x_n)$ and $\epsilon^{(l)} \sim p(\epsilon)$.

At learning stage, we use mini-batch stochastic optimization to optimize the ELBO, suppose our dataset contains N cells and the size of each mini-batch is M, we can get the estimator of marginal likelihood lower bound of the stochastic mini-batch:

$$L(\theta, \varphi; x_n) \cong L^M(\theta, \varphi; x_n) = \frac{1}{M}\sum_{i=1}^{M} \tilde{L}(\theta, \varphi; x_n)$$

When M is large enough, Diederik et al.[59] found that the number of samples L per datapoint can even be set to 1, hence decrease the time consumption when conduct expectation estimation for $E_{q_\varphi(z_n|x_n,s_n)}[\log p_\theta(x_n|z_n, s_n)]$. Throughout our experiment we set M = 128 data points to guarantee the large-sample requirement. We use Adam optimizer with learning rate = 0.01. We also use deterministic warm-up and batch normalization during learning to learn an expressive model which is recommended by Sonderby et al.[60]

Bayesian biclustering incorporate biological information

After reconstructing the original expression matrix, we can conduct biclustering procedure to detect condition-specific FGMs and identify cell subpopulations with distinct functions. Relevant studies have shown that if we can introduce existing biological information (such as the metabolic pathways from the KEGG database) into the process of biclustering, then the accuracy of the biclustering results will be improved.[20,53,61–64] Therefore, we adopt a Bayesian analysis framework, which can introduce prior information to guide variable selection.

Suppose our reconstructed data matrix is $\mathbf{X}$ of size $p \times n$, where p represents the number of genes and n is the number of cells. In order to reduce randomness, here we do not directly decompose the data matrix $\mathbf{X}$, but decompose its parameter matrix. We denote the parameter matrix of $\mathbf{X}$ as μ (e.g., mean) and decompose it: $\mu = \mathbf{m1}^T + \mathbf{WZ}$, where $\mathbf{m}$ is a $p \times 1$ bias vector, $\mathbf{1}$ is a $n \times 1$ vector of 1, $\mathbf{W}$ is a $p \times L$ matrix containing the bicluster information at gene level, indexes of non-zero rows of column l denotes the involved genes in bicluster l. $\mathbf{Z}$ is a $L \times n$ matrix containing the bicluster information at cell level, indexes of non-zero columns of row l denotes the involved cells in bicluster l. Since the observation of gene j from cell i x_{ji} is generated independently, the likelihood function of $\mathbf{X}$ is the product of the likelihood functions of each independent observation. Here we set x_j to be a random variable that follows Gaussian distribution with a likelihood function π_j in the discussion following on.

The likelihood function of an individual observation is:

$$\pi_j(x_{ji}|\mu_{ji}, \rho_j) = \frac{\rho_j^{1/2}}{\sqrt{2\pi}} e^{-\rho_j(x_{ji}-\mu_{ji})^2/2}, x_{ji} = 0, 1..... \quad \text{(Equation 1)}$$

Now we discuss how to introduce prior information. To obtain a sparse estimate of $\mathbf{W}$, we first use the Laplace prior on the matrix $\mathbf{W}$:

$$log\ \pi(\boldsymbol{W}|\lambda) = C + \sum_{j,l} log\ \lambda_{jl} - \sum_{j,l} \lambda_{jl}|w_{jl}|$$

Here the prior parameter λ controls the degree of shrinkage of w. Unlike standard Laplacian prior that uses the same shrinkage parameter λ for all w_{jl}'s, we use different shrinkage parameter for individual w_{jl} to achieve adaptive shrinkage. To incorporate biological information represented by a given graph $\mathcal{G} = <P,E>$, we consider the intuitive scenario where there is an edge between p_1 and p_2, as well as another edge between p_2 and p_3. In this case, if p_1 is selected, we encourage the selection of p_2, and if p_2 is selected, we encourage the selection of p_3. However, if p_1 is selected but p_2 is not, we do not encourage the selection of p_3. To achieve these, we propose encouraging one variable to load on a factor if the other connected variable exhibits a non-zero loading on the same factor. Applying this concept to notations, if x_j and x_k are directly connected in $\mathcal{G}$ and w_{jl} is non-zero for some l, we encourage w_{kl} to also have non-zero values. For this purpose, we introduce a graph-Laplacian prior for λ given the precision matrix Ω as:

$$log\ \pi(\alpha|\Omega) = C_{v_2} + \frac{L}{2} log|\Omega| - \frac{1}{2v_2}\sum_l (\alpha_l - v_1\mathbf{1})\Omega(\alpha_l - v_1\mathbf{1}) \quad \text{(Equation 2)}$$

Where $\alpha_{jl} = \log\lambda_{jl}$, $\alpha_l = (\alpha_{1l}, \ldots, \alpha_{pl})^T$, v_1 and v_2 are hyperparameters. The precision matrix Ω connecting the correlated λ is defined as:

$$\Omega = \begin{bmatrix} 1+\sum_{j\neq 1}\omega_{1j} & -\omega_{12} & \cdots & -\omega_{1p} \\ -\omega_{21} & 1+\sum_{j\neq 2}\omega_{2j} & \ddots & -\omega_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ -\omega_{p1} & -\omega_{p2} & \cdots & 1+\sum_{j\neq p}\omega_{pj} \end{bmatrix}$$

Ω is a symmetric matrix, i.e., $w_{jl} = w_{lj}$. and the prior of Ω is assigned on set $\boldsymbol{\omega} = \{\omega_{jk} : j < k\}$:

$$\pi(\omega) \propto |\Omega|^{-\frac{L}{2}} \prod_{(j,k)\in E} \omega_{jk}^{a_\omega - 1} \exp(-b_\omega\omega_{jk}) 1(\omega_{jk} > 0) \prod_{(j,k)\notin E} \delta_0(\omega_{jk}) \quad \text{(Equation 3)}$$

$\delta_0(\cdot)$ is the Dirac function centered at 0, $1(\cdot)$ is an indicative function. a_ω and b_ω are two hyper-parameters needed to be specified *a priori*. Suppose genes function in similar pathways are connected in a prior graph $\mathcal{G}$, say, if x_j and x_k are directly connected in $\mathcal{G}$, then (3) will try to make the precision matrix components ω_{jk} to be non-zero, and make the contraction term λ_{jl} and λ_{kl} related through (2). In the resulting matrix **W** containing the bicluster information at gene level, since w_{jl} and w_{kl} are subject to a similar degree of contraction under this condition, they tend to be both zero or non-zero at the same time. In other words, if genes j and k are directly connected in similar pathways, they are encouraged to be selected together (or not selected together) in bicluster. Therefore, a standout feature of this approach is that the selected feature set in each bicluster tends to include functional gene module rather than individual genes, resulting in more biologically meaningful results.

Since the **Z** matrix represents the results on the cell set, there is no special pathway information between the samples, so it is sufficient to perform Laplace sparse prior on it:

$$\log\ \pi(\boldsymbol{Z}|\xi) = C + \sum_{l,j}\log\ \xi_{li} - \sum_{l,j}\xi_{li}|z_{li}|$$

Where ξ is the contraction factor, on which a conjugate prior is applied, i.e., Gamma prior:

$$\log\ \pi(\xi) = C_{v_3,v_4} + (v_3 - 1)\sum_{l,j}\log\ \xi_{il} - \frac{1}{v_4}\sum_{l,j}\xi_{li} \quad \text{(Equation 4)}$$

v_3 and v_4 are another two hyper-parameters needed to be specified *a priori*.

Prior specification

In this Bayesian setting, several parameters need to be specified *a priori*, including ν_1 and ν_2 from Equation 2, a_ω and b_ω from Equation 3, and ν_3 and ν_4 from Equation 4. Based on our experience with numerical experiments, we have set a_ω as 4 and b_ω as 1. This choice ensures that the prior correlation for ω is large while maintaining a relatively uninformative prior. Furthermore, we have fixed ν_2 as *ln* 2 and ν_3 as 1 to establish a unit coefficient of variation for the corresponding priors of α and ξ The parameters ν_1 and ν_4 play a crucial role in controlling the sparseness of the solutions for $\boldsymbol{W}$ and $\boldsymbol{Z}$, determining the size of each bicluster. After conducting parameter tuning pre-tests, we recommend setting $\nu_1 = 20$ and $\nu_4 = 7$, which we consistently applied throughout our experiments. We have also designated these values as modifiable default parameters within the model.

MAP estimation for biclustering result

In the optimization stage, we adopt the Pólya-Gamma latent variable proposed by Polson et al.[65] We use the identity formula provided in Polson et al.[65]:

$$\frac{e^{\mu_{ji}x_{ji}}}{(1+e^{\mu_{ji}x_{ji}})^{b_{ji}}} = 2^{-b_{ji}} e^{\kappa_{ji}\mu_{ji}} \int_o^\infty e^{-\rho_{ji}\mu_{ji}^2/2}\pi_{ji}(\rho_{ji})d\rho_{ji}$$

Where $\kappa_{ji} = x_{ji} - b_{ji}/2$, $\pi_{ji}(\rho_{ji})$ is of the Pólya-Gamma class $\mathcal{PG}(b_{ji}, 0)$. So Equation 1 can be written as:

$$\pi_j(\mathbf{x_j}|\mu_j) \propto e^{-\frac{1}{2}\sum_i \rho_{ji}(\mu_{ji}-x_{ji})^2} \pi_j^*(\rho_j)$$

Where $\rho_j \sim \mathcal{G}(\frac{\zeta_j+n}{2}, \frac{\zeta_j}{2})$, ζ_j is the prior parameter for variance. After the introduction of latent variable ρ, LASSO can be efficiently solved in the M step of the EM algorithm. Here we use dynamic weighted LASSO algorithm to speed up the calculation.[66] Additionally, we utilize maximum a posteriori estimation (MAP) to estimate the parameters, which is defined as:

$$(\widehat{\mathbf{W}}, \widehat{\mathbf{Z}}, \widehat{\mathbf{m}}, \widehat{\alpha}, \widehat{\xi}) = \mathrm{argmax}_{\mathrm{W,Z,m},\alpha,\xi} \int\int \pi(\mathbf{W}, \mathbf{Z}, \mathbf{m}, \boldsymbol{\alpha}, \xi, \boldsymbol{\rho}, \boldsymbol{\Omega}|\mathbf{X}) \mathrm{d}\rho \mathrm{d}\Omega$$

This can be efficiently solved using the EM algorithm, and the objective function at t iterations is:

$$\boldsymbol{Q}_t(\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{m}, \alpha, \xi) = -\frac{1}{2}\sum_{i,j}\rho_j^{(t)}(\mu_{ji} - x_{ji})^2 + \sum_{j,l}\alpha_{jl} - \sum_{j,l}\lambda_{jl}|w_{jl}| + v_3\sum_{l,i}\log\xi_{l,i} - \sum_{l,i}\xi_{l,i}\left(|z_{li}| + \frac{1}{4}\right) - \frac{1}{2v_2}\sum_l(\alpha_l - v_1\mathbf{1})^T\Omega^{(t)}(\alpha_l - v_1\mathbf{1})$$

Where $\mu = \boldsymbol{m}^{(t-1)} + \boldsymbol{W}^{(t-1)}\boldsymbol{Z}^{(t-1)}$, $\rho_{ij}^{(t)} = E(\rho_{ij}|\boldsymbol{X}, \boldsymbol{W}^{(t-1)}, \boldsymbol{Z}^{(t-1)}, \boldsymbol{m}^{(t-1)}, \alpha^{(t-1)}, \xi^{(t-1)})$ and $\Omega^{(t)} = E(\omega_{ij}|\boldsymbol{X}, \boldsymbol{W}^{(t-1)}, \boldsymbol{Z}^{(t-1)}, \boldsymbol{m}^{(t-1)}, \alpha^{(t-1)}, \xi^{(t-1)})$.

Strong classification of cell group for different biclustering methods

For scBC and GBC, we can directly observe the contribution of each bicluster to the parameter matrix in each cell from the result of the Z matrix. We assign a cell to the most involved cluster, which is determined by the row with the largest absolute value. For the remaining methods, we aim to achieve optimal cell classification results without losing any information from the biclustering results. Due to the high degree of cell overlap in the biclustering results, we convert the cell-level biclustering results into a graph, where cells in the same bicluster are connected by edges. If the occurrences of a pair of cells increase, the weights of the edges between them also increase accordingly. We then apply the Markov clustering algorithm (MCL)[26] to convert the graph into cell-level clustering results. For each method, we set the number of iterations to a value between 1 and 10 that allows the method to achieve the highest adjusted Rand index (ARI). In fact, we observed that the number of iterations required for the best results usually does not exceed 7. After the transformation, each cell is exclusively assigned to a cluster, and we can evaluate the cell-level clustering results using any clustering evaluation criterion.

Datasets and preprocessing

Here we describe all of the datasets and the preprocessing steps used in the current work as follows. The prior information for all the real-world datasets is extracted by *biomaRt* using the highly variable genes.

HEART

This is a combined single cell and single nuclei RNA-Seq data of 485K cardiac cells with annotation from Heart Cell Atlas. Here we use a subsampled version provided at https://github.com/YosefLab/scVI-data/blob/master/hca_subsampled_20k.h5ad, which has been filtered down randomly to 20k cells. In our study, we further filtered 1000 highly variable genes using *scanpy* and generate 10 subsampled datasets with each containing 1000 randomly selected cells.

PBMC

This actually is a purified PBMC dataset from.[50] An organized version can be accessed from https://github.com/YosefLab/scVI-data/raw/master/PurifiedPBMCDataset.h5ad. We also conducted a subsampling procedure here: first screen out 853 highly variable genes using *scanpy*, then generate 10 subsampled datasets with each containing 1000 random selected cells.

LUAD

Single cell RNA sequencing of lung adenocarcinoma from,[51] which can be accessed from the NCBI Expression Omnibus database (accession code GSE131907). This is single cell RNA sequencing (scRNA-seq) for 208,506 cells derived from 58 lung adenocarcinomas from 44 patients, which covers primary tumor, lymph node and brain metastases, and pleural effusion in addition to normal lung tissues and lymph nodes. Here we use *Seurat* to conduct preprocessing: we first randomly selecet 10000 cells to filter 2000 highly variable genes, then generate 10 subsampled datasets with each containing 5000 cells.

BC

Single cell RNA sequencing of primary breast cancer from,[52] which can be accessed from the NCBI Expression Omnibus database (accession code GSE75688). This dataset contains 515 cells from 11 patients and most of the cell type is tumor. We first screen out 2000 highly variable genes using *Seurat* then conduct subsampling. Due to the serious category imbalance problem in this dataset (326 cells are labeled as "Tumor"), we only sample 86 cells with the tumor label each time and all cells with other labels are retained so that the results will not be unreliable due to class imbalance during evaluation.

AD

A total of 80660 droplet-based single-nucleus RNA-seq (snRNA-seq) profiles for Alzheimer's disease from.[3] The postmortem human brain samples came from 48 participants in the Religious Order Study (ROS) or the Rush Memory and Aging Project (MAP), collectively known as ROSMAP with 24 individuals with high levels of β-amyloid and other pathological hallmarks of AD ('AD-pathology'),

and 24 individuals with no or very low β-amyloid burden or other pathologies ('no-pathology'). The original study clustered individuals based on nine clinico-pathological traits to further define the pathology groups as 'early-pathology' and 'late-pathology'. And that division is totally adopted in our study. The snRNA-seq data are available on The Rush Alzheimer's Disease Center (RADC) Research Resource Sharing Hub at https://www.radc.rush.edu/docs/omics.html (snRNA-seq PFC) or at Synapse (https://www.synapse.org/#!Synapse:syn18485175) under the https://doi.org/10.7303/syn18485175. The data are available under controlled use conditions set by human privacy regulations. To access the data, a data use agreement is needed. Since we are not going to use this dataset to conduct benchmarking here, there is no need to repeatedly generate subsamples. When preprocessing the dataset, we first use stratified sampling to draw one out of ten cells, then 2000 highly variable genes are refined by Seurat. This sample is then used to be explored later.

Simulated data

In each simulation setting, we generate 100 simulated datasets. For convenience, we denote p as the number of genes, n as the number of cells. The scale of the FGM increases adaptively with the size of the simulated dataset (actually the size of p). The parameter μ is computed by the multiplicative model $\mu = WZ$, where W is a p×L matrix and Z is an L×n matrix. The number of non-zero elements in each column of W is set as p/20, and the number of non-zero elements in each row of Z is randomly drawn from a Poisson distribution with a parameter of 30. The row indices of non-zero elements in W and the column indices of Z with non-zero elements are randomly drawn from 1 to p and 1 to n. The nonzero element values for both W and Z are generated from a normal distribution with mean 1.5 and standard deviation 0.1, and are randomly assigned to be positive or negative. The prior edge is generated along with W.

When generating X, each element is generated from $\mathrm{NB}\left(r_j, \frac{1}{1+e^{-\mu_{ij}}}\right)$, and the parameter r_j is randomly drawn from 5 to 20. Finally, in order to simulate different batches, we divided the dataset into three parts, each of n/3 samples, with different intensities of noise. The implementation of dropout is to perform Bernoulli censoring at each data point according to the given dropout rate parameter.

The simulation data generation process is shown in Figure 3A.

Matching of biclusters when analyzing AD dataset

To avoid confusion, we first explain the difference between biclustering and bicluster, two concepts we've been using throughout the paper. A biclustering refers to execute one biclustering algorithm once (e.g., scBC). After biclustering is conducted, we can get several columns-rows pairs, each is called a bicluster. In our study, we conduct three independent biclusterings on the three pathologically seperated AD datasets, each with L biclusters. L is the number of biclusters we set beforehand.

When conducting scBC on the AD dataset, genes that are widely present in all biclusters represent the commonality among all cells. We subtract these genes from each bicluster to reduce the homogeneity of different biclusters, since we are not interested in them in this case. Here, we set the number of biclusters in each round of biclustering to 6. However, this is an empirical hyperparameter, and some biclusters may have a high degree of similarity and be more reasonable to merge into a single bicluster. This applies not only to different biclusters in a whole biclustering but also to different biclusters in independent biclusterings. However, the methods for aligning biclusters in a single biclustering and for biclusters in different biclusterings should be different, since biclusters from the latter are somewhat more independent.

Due to their own homogeneity or correlation, more attention should be paid to the exclusivity when merging biclusters from a single biclustering. We denote the number of genes only appear in biclusters i and j as e, which means all the other biclusters don't have these genes. And genes present in bicluster k is denoted as g_k, the overlap score is defined as:

$$os_{i,j} = \frac{e}{\max\{|g_i|, |g_j|\}}$$

In our study, a pair of biclusters with overlap score >0.03 are combined as an FGM. The biclusters correspondences in different biclusterings are independent, so more attention should be paid to the degree of overlap. Here we use the ''overlap over union''(IoU) criterion to combine different biclusters:

$$IoU_{i,j} = \frac{intersect\{g_i, g_j\}}{union\{g_i, g_j\}}$$

each pair of biclusters with IoU >0.3 are combined as an FGM. The alignment results are in Table S14.

FGM perturbation during AD progression and enrichment analysis

Before merging functional gene modules (FGMs) from different biclusters, we first assign each cell exclusively to a bicluster using the strong classification method as before. When similar FGMs are combined, functionally related cells are also merged as a whole. Next, we obtain the FGM composition contained in each cell type. As observed, multiple FGMs can be simultaneously active in one cell type. To illustrate FGM perturbation for each cell type during AD progression, we first observe the changes in the proportion of each FGM in each cell type. FGMs with an elevated ratio are candidates for "increased activity," while those with a reduced ratio are candidates for "decreased activity." We then examine the differences in the gene set makeup of these two types of FGMs. The overlap between the two represents commonalities exhibited in certain cell types, which are not of interest to us. We focus on the exclusive genes of "increased activity" and "decreased activity," which may uncover pathway perturbations in different

pathological states. The exclusive genes in "increased activity" and "decreased activity" are used for functional enrichment using *clusterProfiler*. The results are used to reveal the pathway perturbation during AD progression.

Analyzing AD dataset using GBC

To ensure a fair comparison, we employed the identical analysis pipeline for the AD dataset as used by scBC for GBC. The only distinction lies in the threshold for the overlap score when matching different biclusters within a single biclustering. In this case, a threshold of 0.2 was set to avoid merging all biclusters into a single FGM, as it would yield inconclusive results. The matching results of GBC are in Table S15.

QUANTIFICATION AND STATISTICAL ANALYSIS

Here we will describe the evaluation metrics used in our study.

Comparison of FGM detection

To quantify the performance of each method in detecting functional gene modules (FGMs), we conduct gene ontology (GO) enrichment using *clusterProfiler* for each gene set of each bicluster and record the most significant p value (BH adjusted). Since the number of biclusters detected by each method differs, we take the most significant p value of all the biclusters detected by a single method and transform it using -log10(p) to denote the performance of this method. Methods that fail to detect any bicluster are labeled as 0. We use 10 subsamples from each dataset for repeated evaluation.

Criterions for clustering performance

There are three metrices we used to benchmark the clustering performance at cell level: ARI, FMI and AMI. Here we will briefly describe how to compute these metrices:

ARI. The Rand Index computes a similarity measure between two clusterings by considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted and true clusterings. The raw RI score is then "adjusted for chance" into the ARI score using the following scheme:

$$ARI = \frac{RI - E(RI)}{max(RI) - E(RI)}$$

To calculate this value, first calculate the contingency table like that:

	Y_1	Y_2	...	Y_s	Sums
X_1	n_{11}	n_{12}	$\cdots$	n_{1s}	a_1
X_2	n_{21}	n_{22}	$\cdots$	n_{2s}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_r	n_{r1}	n_{r1}	$\cdots$	n_{rs}	a_r
Sums	b_1	b_2	$\cdots$	b_s	

each value in the table represents the number of data point located in both cluster (Y) and true class (X), and then calculate the ARI value through this table:

$$\overbrace{ARI}^{\text{Adjust Index}} = \frac{\overbrace{\sum_{ij}\binom{n_{ij}}{2}}^{\text{Index}} - \overbrace{\left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] \Big/ \binom{n}{2}}^{\text{Expected Index}}}{\underbrace{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right]}_{\text{Max Index}} - \underbrace{\left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] \Big/ \binom{n}{2}}_{\text{Expected Index}}}$$

The adjusted Rand index is thus ensured to have a value close to 0.0 for random labeling independently of the number of clusters and samples and exactly 1.0 when the clusterings are identical (up to a permutation). The adjusted Rand index is bounded below by −0.5 for especially discordant clusterings.

FMI. The Fowlkes-Mallows index (FMI) is defined as the geometric mean between of the precision and recall:

$$FMI = \frac{TP}{\sqrt{(TP+FP)(TP+FN)}}$$

Where TP is the number of **True Positive** (i.e., the number of pair of points that belongs in the same clusters in both true labels and predicted labels), FP is the number of **False Positive** (i.e., the number of pair of points that belongs in the same clusters in true labels but not in predicted labels) and FN is the number of **False Negative** (i.e., the number of pair of points that belongs in the same clusters in predicted labels but not in true labels). The score ranges from 0 to 1. A high value indicates a good similarity between two clusters.

AMI. The Mutual Information is a measure of the similarity between two labels of the same data. Where $|U_i|$ is the number of the samples in cluster U_i and $|V_j|$ is the number of the samples in cluster V_j, the Mutual Information between clusterings U and V is given as:

$$MI(U,V) = \sum_{i=1}^{|U|}\sum_{j=1}^{|V|}\frac{|U_i \cap V_j|}{N}\log\frac{N|U_i \cap V_j|}{|U_i||V_j|}$$

Adjusted Mutual Information (AMI) is an adjustment of the Mutual Information (MI) score to account for chance. It accounts for the fact that the MI is generally higher for two clusterings with a larger number of clusters, regardless of whether there is actually more information shared. For two clusterings U and V, the AMI is given as:

$$AMI(U,V) = \frac{MI(U,V) - E(MI(U,V))}{\mathrm{avg}(H(U),H(V)) - E(MI(U,V))}$$

Where $H(*)$ is the information entropy for a label's distribution (e.g., $H(U) = \sum_{i=1}^{|U|} P(i)log(P(i))$). This metric is independent of the absolute values of the labels: a permutation of the class or cluster label values won't change the score value in any way.

Metrices for biclustering comparison

Suppose M: {1 … L} → {1 … L} maps the ground true bicluster index to the index of the bicluster detected by an algorithm, T_i denote the i_{th} ground true bicluster and B_i denote the i_{th} detected bicluster. The Cluster Error (CE) proposed by Anne et al.[67] is defined as:

$$1 - \mathrm{CE}(M) = \frac{\sum_{i=1}^{L}|T_i \cap B_{M(i)}|}{\left|\bigcup_{i=1}^{L} T_i \cup B_{M(i)}\right|}$$

This is a distance measure of subspace clustering with lower CE indicating better consistency with ground truth. When we evaluate the performance, we choose an *M* minimizing the CE as the optimal match and is used by other measurements. The corresponding 1-CE is output with the higher the value, the better.

We also use F-score (F) to evaluate the performance. F-score is the harmonic mean of precision (PRE) and recall (REC). Here we use the calculation way proposed by Zhong et al.[18]:

$$\mathrm{PRE}_i = \frac{|T_i \cap B_{M(i)}|}{|B_{M(i)}|}$$

$$\mathrm{REC}_i = \frac{|T_i \cap B_{M(i)}|}{|T_i|}$$

Where A denote all the elements of the expression data. PRE_i and REC_i are computed for bicluster pair i, we finally output the average for each criterion as PRE and REC, along with their harmonic mean as F-score(F), which is a combination of the two, and we also pay more attention to it. The higher this indicator is, the better.

Cell Reports Methods, Volume *4*

Supplemental information

Single-cell biclustering for cell-specific transcriptomic perturbation detection in AD progression

Yuqiao Gong, Jingsi Xu, Maoying Wu, Ruitian Gao, Jianle Sun, Zhangsheng Yu, and Yue Zhang

Supplementary Information

Supplementary Figures

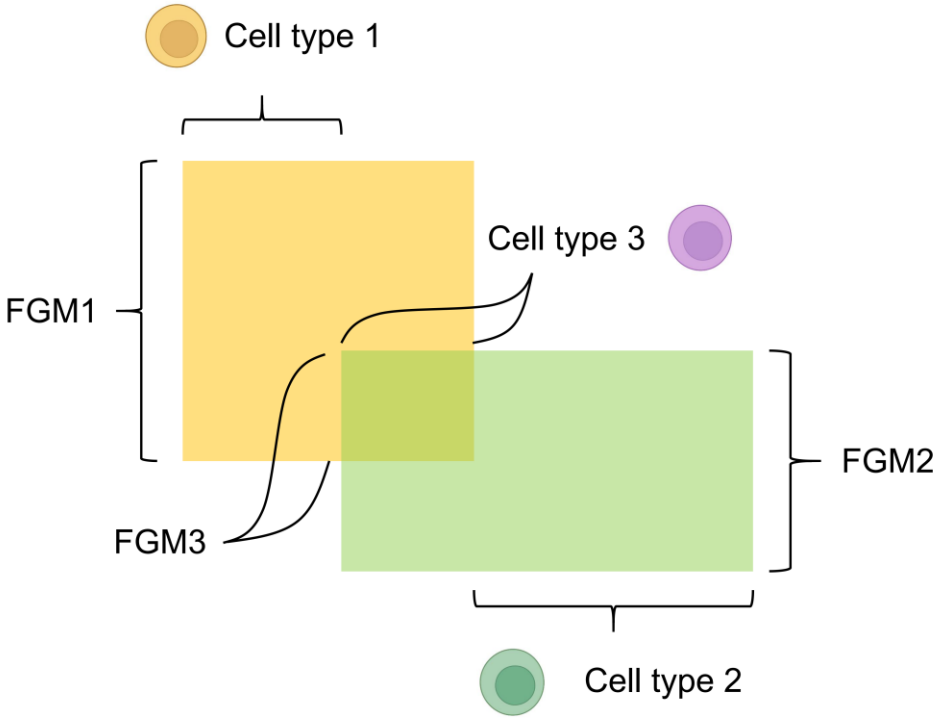


Figure S1. overlap of biclusters, related to Figure 1. This schematic diagram illustrate how biclustering with overlap can uncover the complex patterns in single-cell data. In the complex cell machinery, multiple FGMs are active in a cell group, which is represented by the Figureure that FGM1 and FGM2 are simultaneously active in cell type 3. Meanwhile, different cell groups may share a common FGM. This can be seen from the fact that cell type 1 and cell type 2 share a common FGM3, since FHM3 is contained in both FGM1 and FGM2.

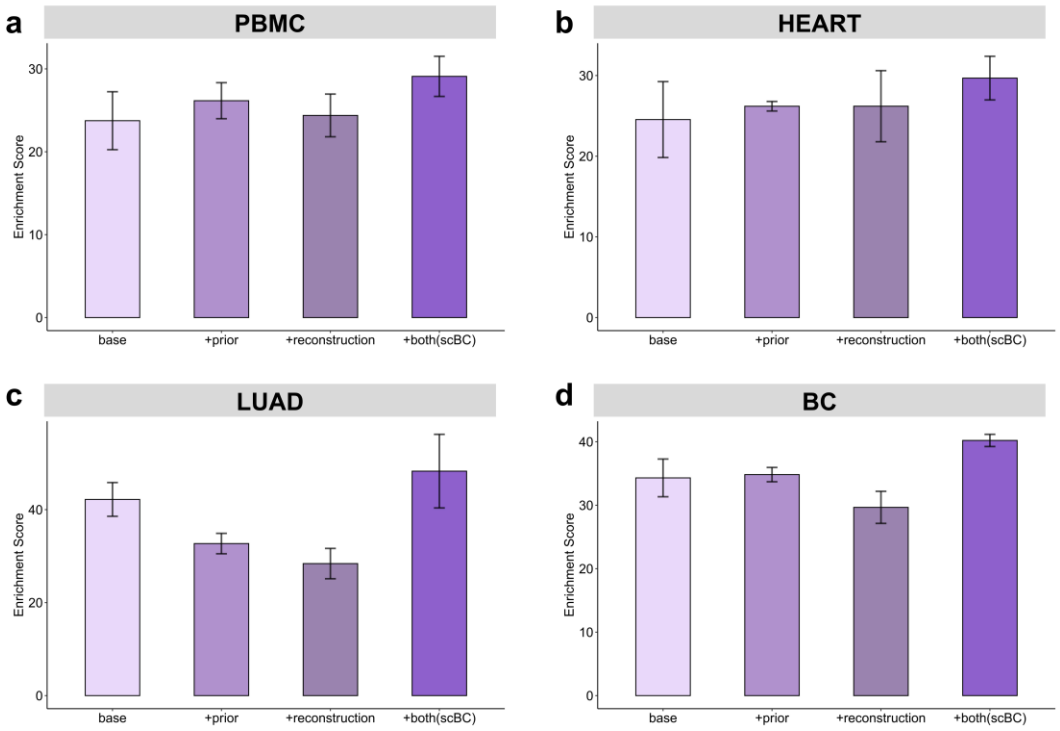


Figure S2. ablation study on four real-world datasets, related to Figure 2. a, Enrichment score of different combinations of building blocks of scBC in PBMC dataset. The x-axis represents different combination strategies: only matrix factorization based biclustering (base); biclustering with prior information (+prior); data reconstruction and biclustering (+reconstruction) and full model (scBC). y-axis represents the enrichment score (-log10(p), BH adjusted). Error bar stands for standard deviation of 10 subsample's results. The notation of the axis is the same for b, c, and d. **b,** Enrichment score of different combinations of building blocks of scBC in HEART dataset. **c,** Enrichment score of different combinations of building blocks of scBC in LUAD dataset. **d,** Enrichment score of different combinations of building blocks of scBC in BC dataset.

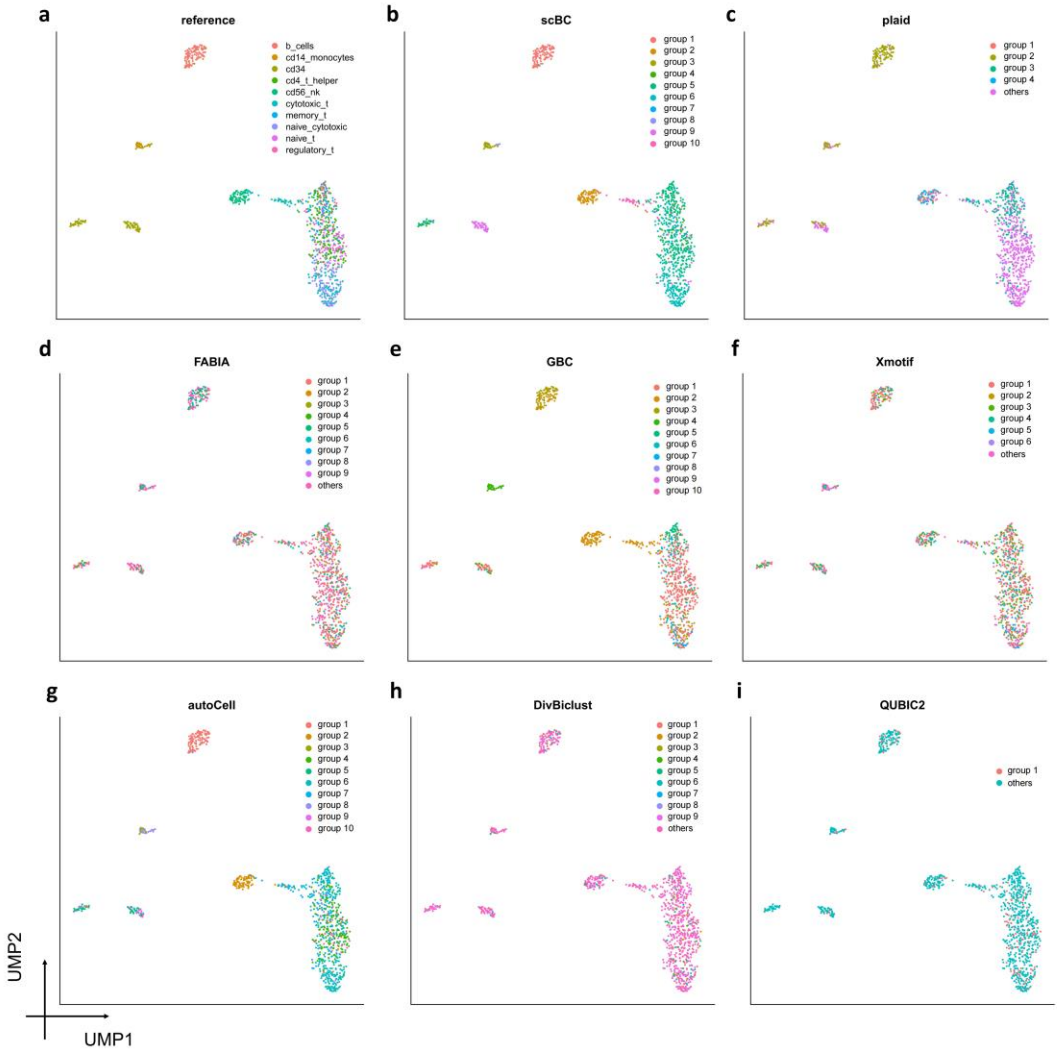


Figure S3. visualization of cell clustering results on PBMC dataset, related to Figure 2f. These are the cell representation of the subsample from the last iteration of PBMC dataset. Some methods output too many categories so that we merge some into "others" whose number of cells is less than one percent of the total sample. Methods not shown here fail to detect functional cell groups.

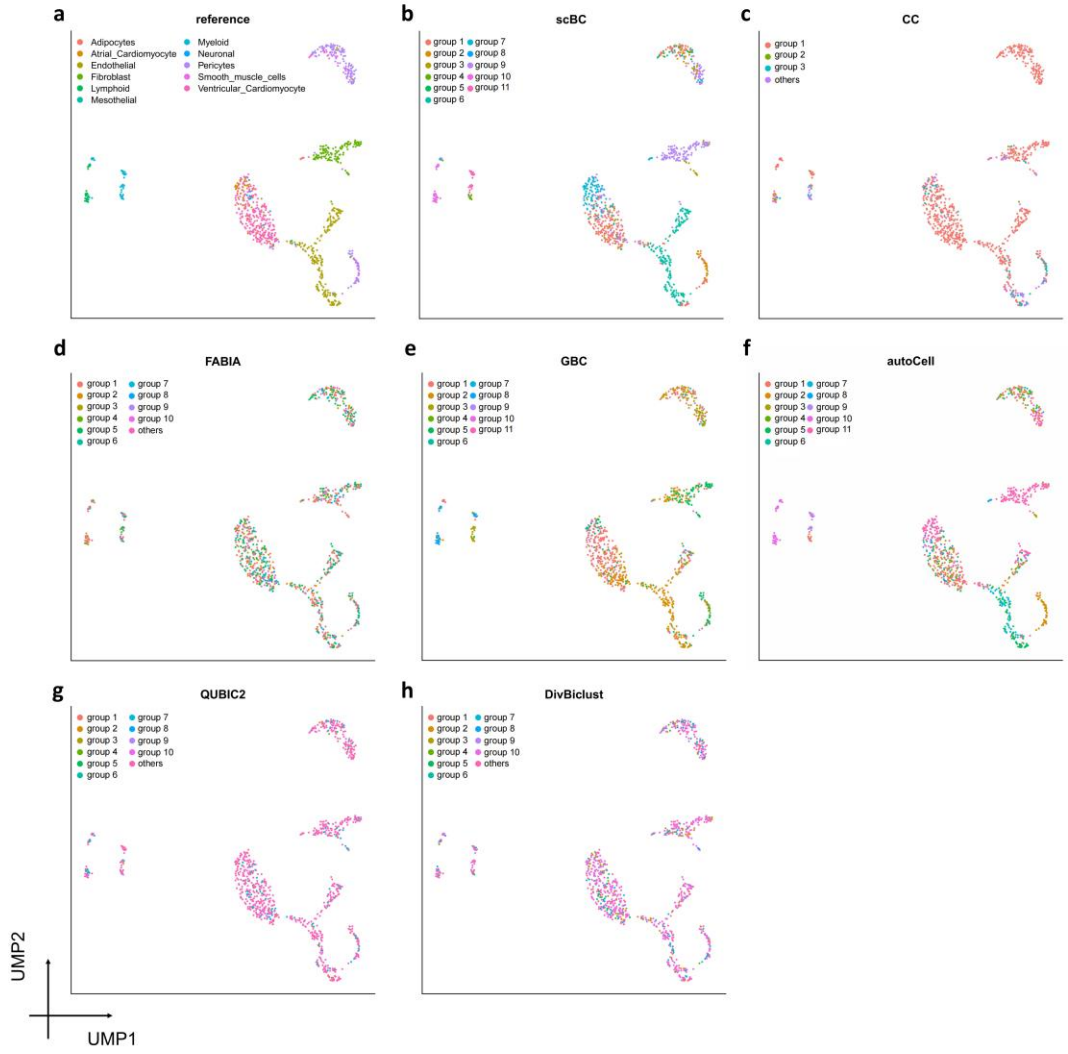


Figure S4. visualization of cell clustering results on HEART dataset, related to Figure 2g. These are the cell representation of the subsample from the last iteration of HEART dataset. Some methods output too many categories so that we merge some into "others" whose number of cells is less than one percent of the total sample. Methods not shown here fail to detect functional cell groups.

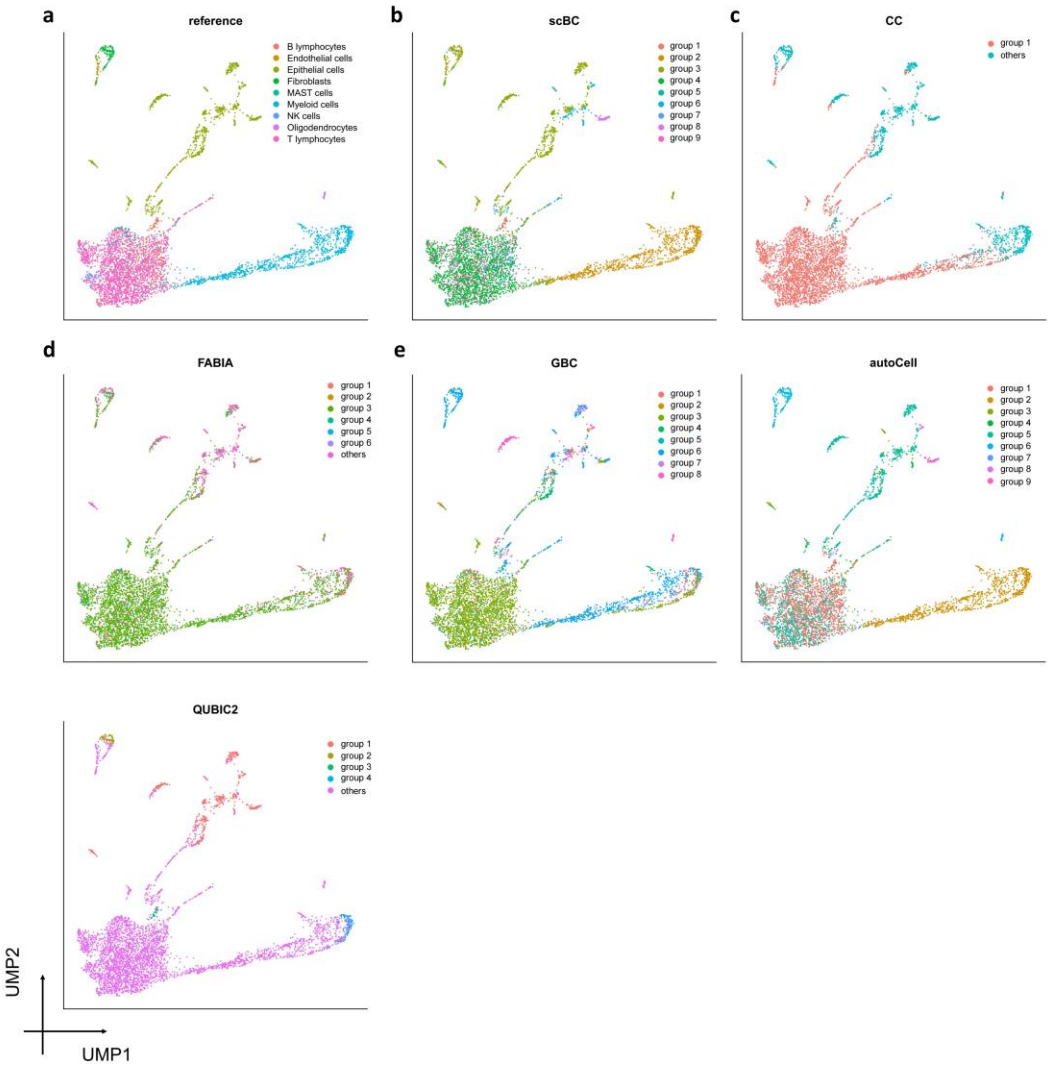


Figure S5. visualization of cell clustering results on LUAD dataset, related to Figure 2h. These are the cell representation of the subsample from the last iteration of LUAD dataset. Some methods output too many categories so that we merge some into "others" whose number of cells is less than one percent of the total sample. Methods not shown here fail to detect functional cell groups.

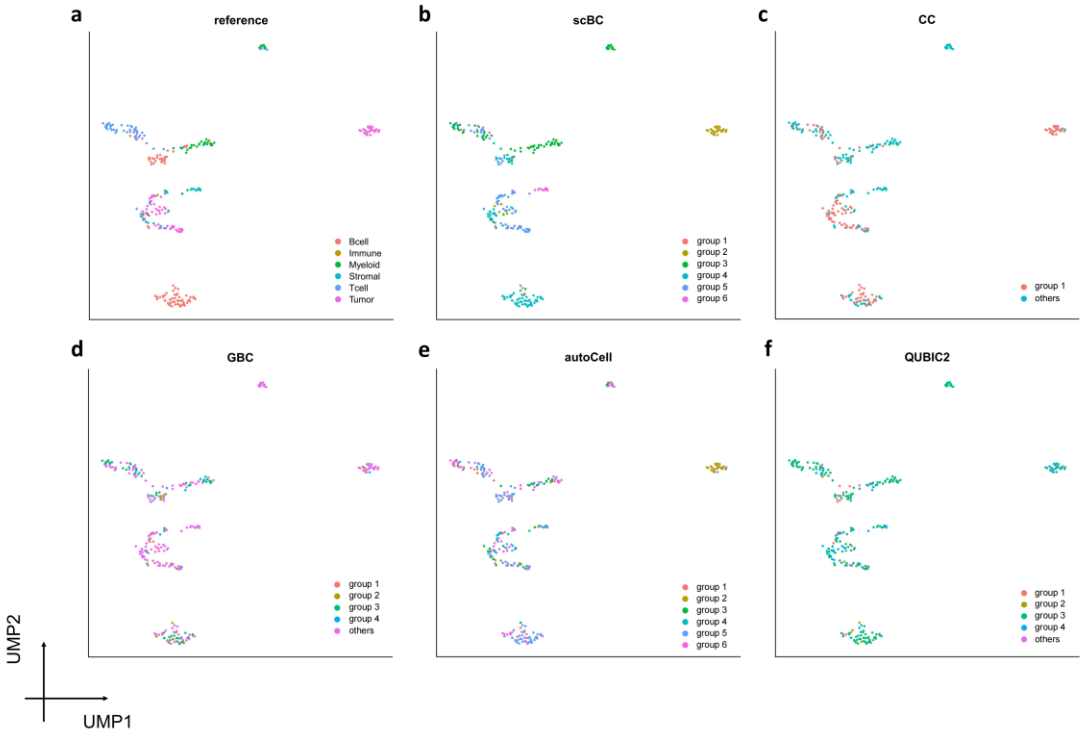


Figure S6. visualization of cell clustering results on BC dataset, related to Figure 2i. These are the cell representation of the subsample from the last iteration of BC dataset. Some methods output too many categories so that we merge some into "others" whose number of cells is less than one percent of the total sample. Methods not shown here fail to detect functional cell groups.

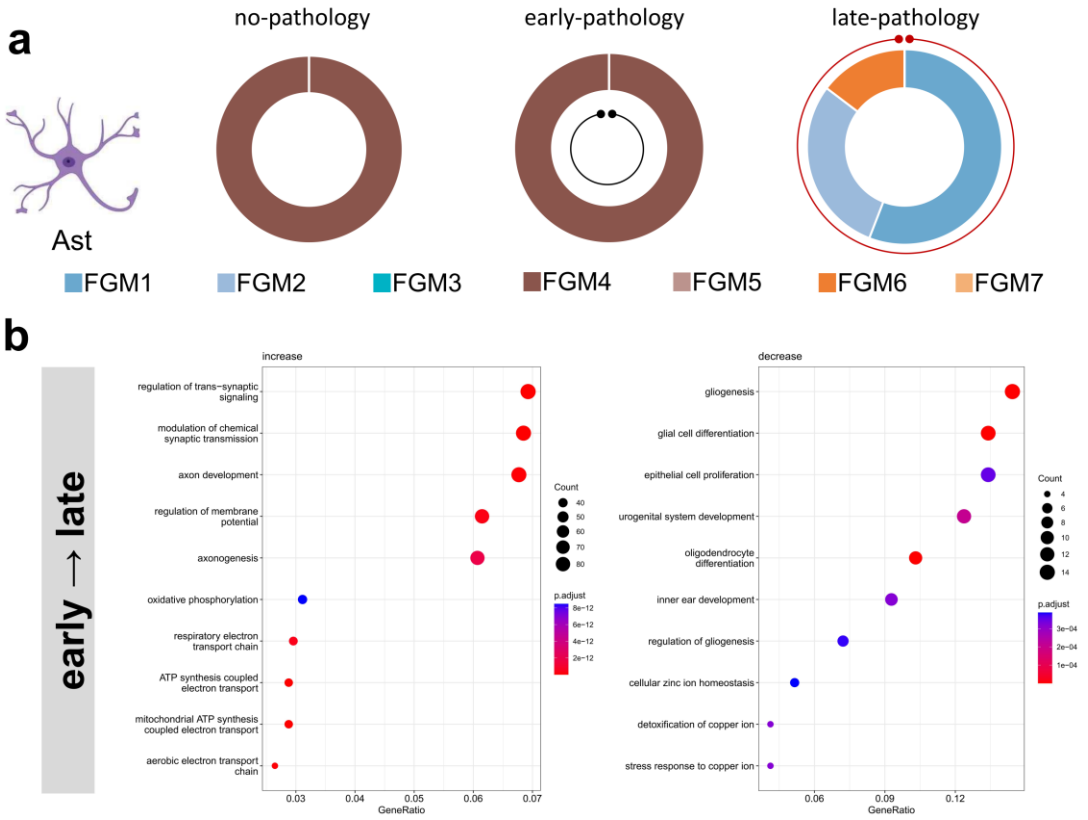


Figure S7. GBC's results of astrocytes on AD dataset, related to STAR Methods. a, perturbation of FGM composition in astrocytes during AD progression. Each pie chart quantifies the FGM composition of a cell under a specific progression condition. The outer red circles indicate FGMs whose composition is increasing compared to the previous stage; The inner black circle represents FGMs whose composition is decreasing compared to the latter stage. **b,** results of enrichment analysis of FGMs altered from early pathology to late pathology. 'increase' represents the specific FGM of the late pathology, and 'decrease' represents the specific FGM of the early pathology, both representing the set of genes that are perturbated during the progression.

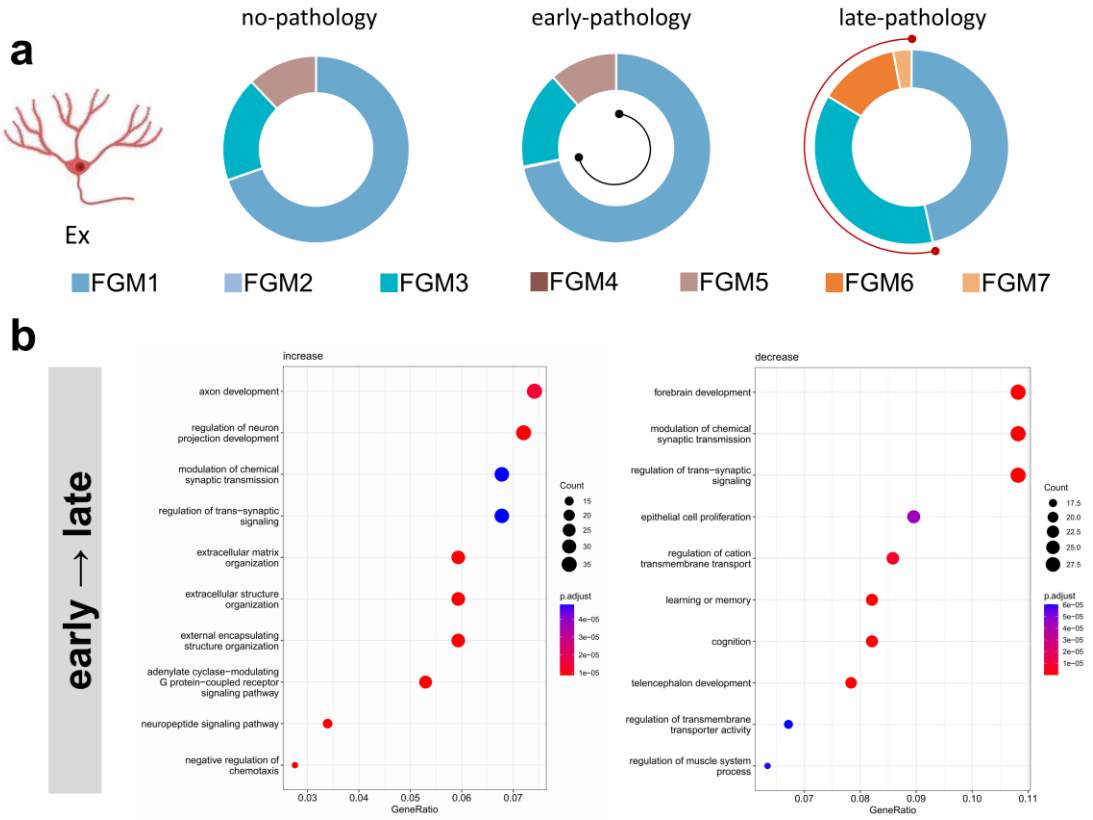


Figure S8. GBC's results of excitatory neurons on AD dataset, related to STAR Methods. a, perturbation of FGM composition in excitatory neurons during AD progression. Each pie chart quantifies the FGM composition of a cell under a specific progression condition. The outer red circles indicate FGMs whose composition is increasing compared to the previous stage; The inner black circle represents FGMs whose composition is decreasing compared to the latter stage. **b,** results of enrichment analysis of FGMs altered from early pathology to late pathology. 'increase' represents the specific FGM of the late pathology, and 'decrease' represents the specific FGM of the early pathology, both representing the set of genes that are perturbated during the progression.

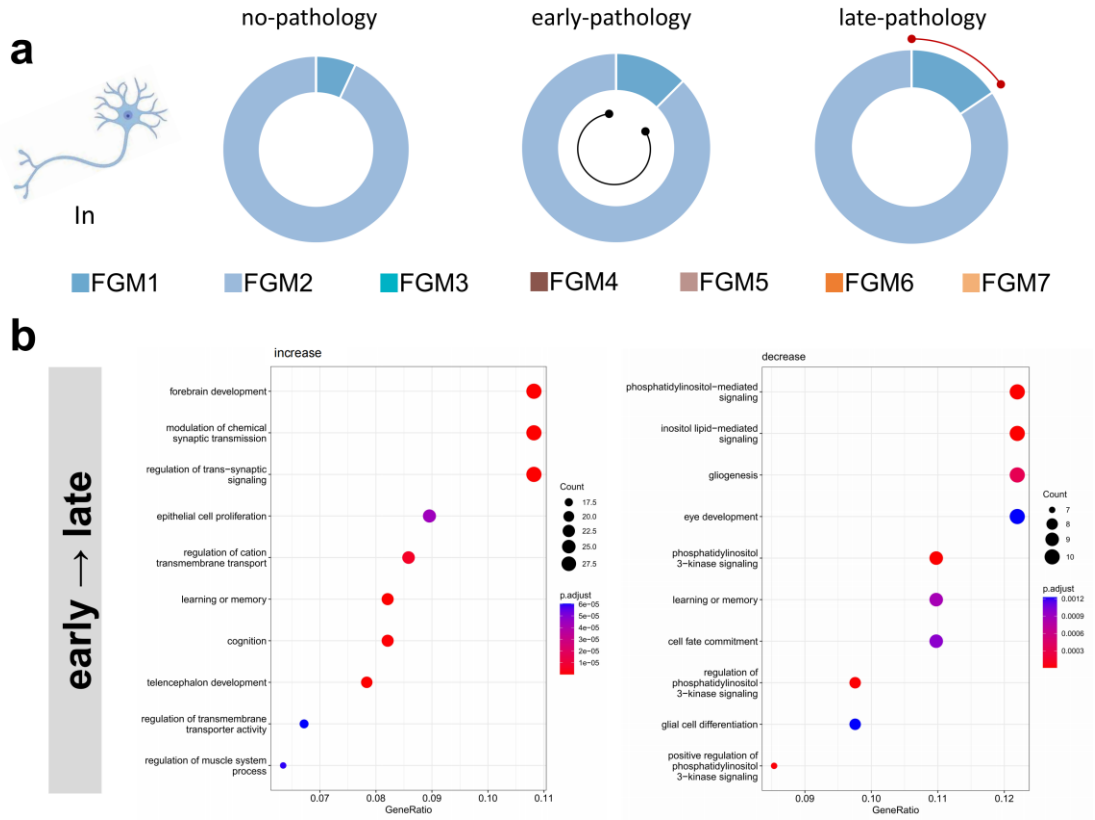


Figure S9. GBC's results of Inhibitory neurons on AD dataset, related to STAR Methods. a, perturbation of FGM composition in Inhibitory neurons during AD progression. Each pie chart quantifies the FGM composition of a cell under a specific progression condition. The outer red circles indicate FGMs whose composition is increasing compared to the previous stage; The inner black circle represents FGMs whose composition is decreasing compared to the latter stage. **b,** results of enrichment analysis of FGMs altered from early pathology to late pathology. 'increase' represents the specific FGM of the late pathology, and 'decrease' represents the specific FGM of the early pathology, both representing the set of genes that are perturbated during the progression.

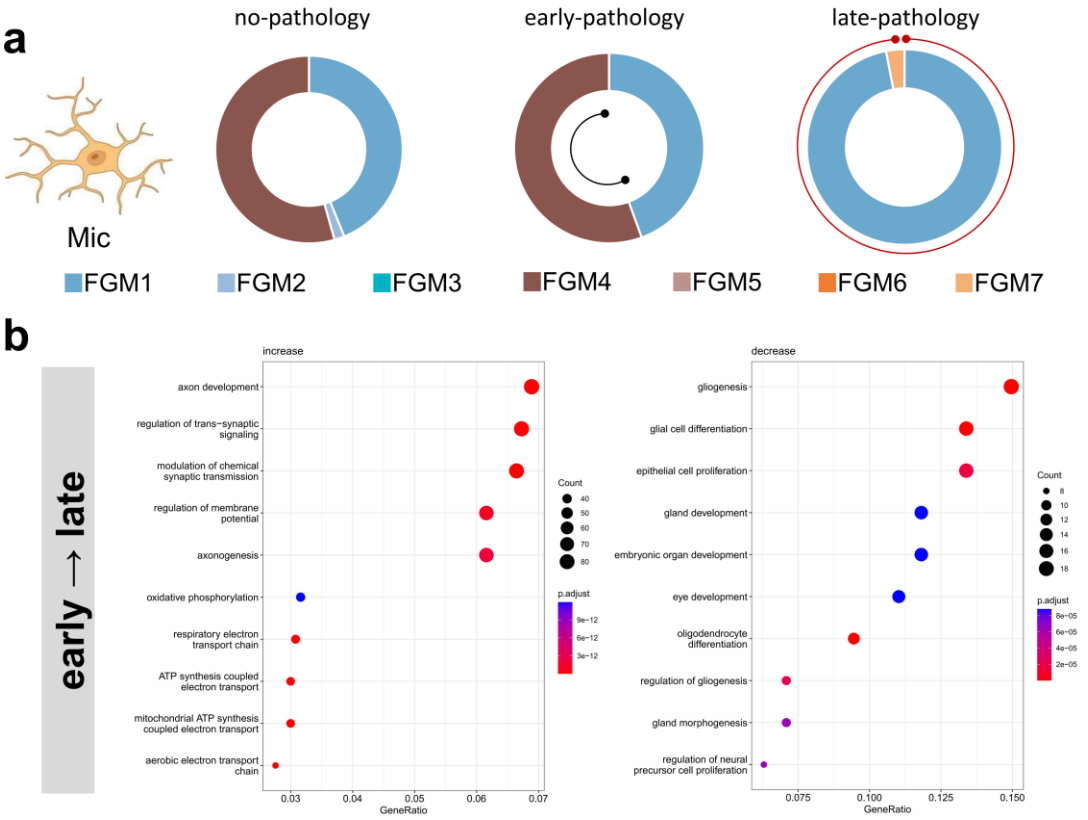


Figure S10. GBC's results of microglia on AD dataset, related to STAR Methods. a, perturbation of FGM composition in microglia during AD progression. Each pie chart quantifies the FGM composition of a cell under a specific progression condition. The outer red circles indicate FGMs whose composition is increasing compared to the previous stage; The inner black circle represents FGMs whose composition is decreasing compared to the latter stage. **b,** results of enrichment analysis of FGMs altered from early pathology to late pathology. 'increase' represents the specific FGM of the late pathology, and 'decrease' represents the specific FGM of the early pathology, both representing the set of genes that are perturbated during the progression.

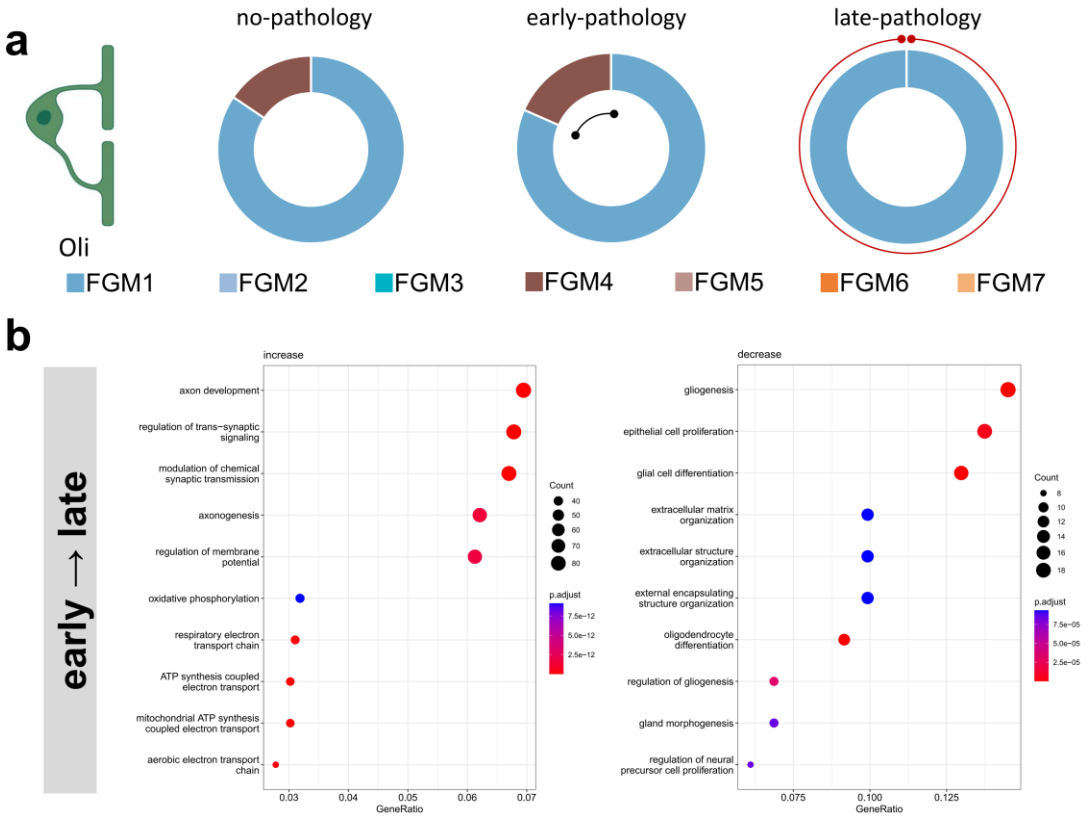


Figure S11. GBC's results of oligodendrocytes on AD dataset, related to STAR Methods. a, perturbation of FGM composition in oligodendrocytes during AD progression. Each pie chart quantifies the FGM composition of a cell under a specific progression condition. The outer red circles indicate FGMs whose composition is increasing compared to the previous stage; The inner black circle represents FGMs whose composition is decreasing compared to the latter stage. **b,** results of enrichment analysis of FGMs altered from early pathology to late pathology. 'increase' represents the specific FGM of the late pathology, and 'decrease' represents the specific FGM of the early pathology, both representing the set of genes that are perturbated during the progression.

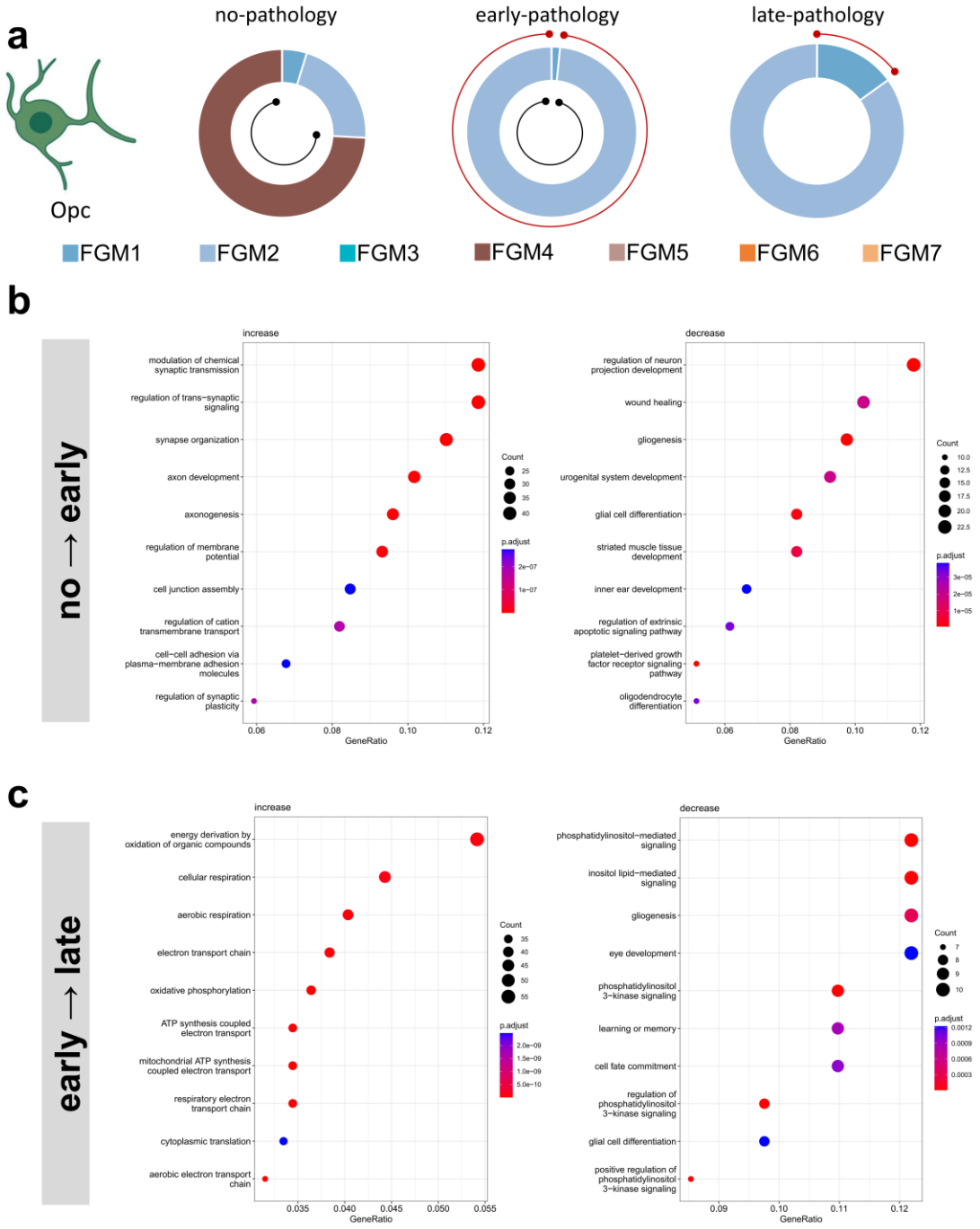


Figure S12. GBC's results of oligodendrocyte precursor cells on AD dataset, related to STAR Methods. a, perturbation of FGM composition in oligodendrocyte precursor cells during AD progression. Each pie chart quantifies the FGM composition of a cell under a specific progression condition. The outer red circles indicate FGMs whose composition is increasing compared to the previous stage; The inner black circle represents FGMs whose composition is decreasing compared to the latter stage. **c,** results of enrichment analysis of FGMs altered from no pathology to early pathology. 'increase' represents the specific FGM of early pathology group, and 'decrease' represents the specific FGM of no pathology group, both representing the set of genes that are perturbated during the progression. The notation is

the same in c. **c,** results of enrichment analysis of FGMs altered from early pathology to late pathology.

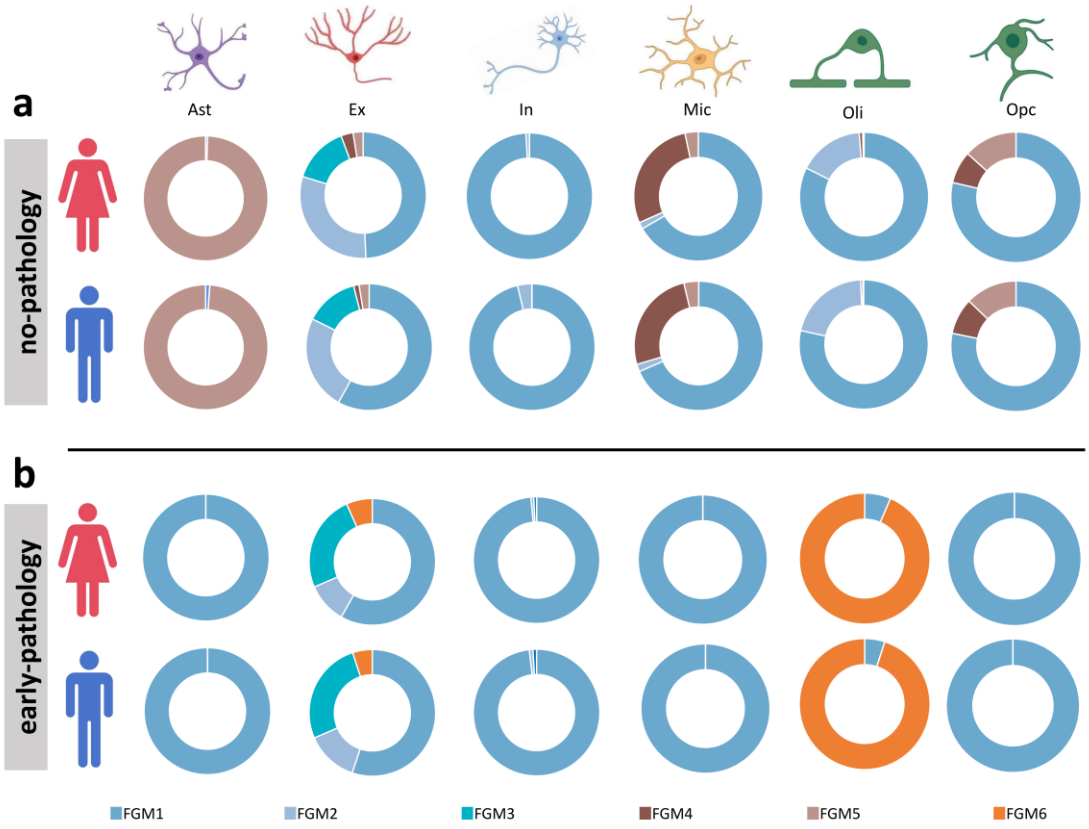


Figure S13. AD data's analyzing results stratified by sex, related to Figure 6. a, FGM composition in each cell type during AD progression in no-pathology group, stratified by sex. **b,** FGM composition in each cell type during AD progression in early-pathology group, stratified by sex.

Supplementary Tables

Table S1. Summarization of different datasets used in our study, related to STAR Methods

Datasets	sc/snRNA-seq	Sequencing protocol	Source
HEART	combined single cell and single nuclei RNA-Seq data	droplet-based	https://github.com/YosefLab/scVI-data/blob/master/hca_subsampled_20k.h5ad
PBMC	scRNA-seq	droplet-based	https://github.com/YosefLab/scVI-data/raw/master/PurifiedPBMCDataset.h5ad
LUAD	scRNA-seq	droplet-based	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE131907
BC	scRNA-seq	plate-based	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE75688
AD	snRNA-seq	droplet-based	https://www.synapse.org/#!Synapse:syn18485175

Table S2 shows the enrichment scores obtained by comparing different methods through repeated experiments (10 repetitions) on different datasets, corresponding to the performance of different methods in discovering functional gene modules (as shown in Figure. 2b of the main text).

Table S2. Enrichment score comparison of different methods, related to Figure 2

datasets / methods	PBMC		HEART		LUAD		BC	
	mean	sd	mean	sd	mean	sd	mean	sd
GBC	23.75	3.49	24.54	4.70	42.20	3.63	34.31	2.97
plaid	9.19	2.63	0.48	1.53	0.74	1.57	0.00	0.00
Xmotif	22.59	1.70	0.00	0.00	0.00	0.00	0.00	0.00
CC	26.86	0.00	23.70	1.55	43.65	1.02	7.46	2.84
Bimax	5.43	3.266	0.00	0.00	0.00	0.00	0.00	0.00
FABIA	21.83	2.72	11.70	2.44	22.37	2.05	0.00	0.00
autoCell	26.50	4.16	29.41	4.03	27.81	2.13	25.24	11.56
QUBIC2	4.02	0.96	4.29	0.83	9.19	3.56	23.09	1.36
scBC	29.10	2.42	29.68	2.70	48.28	7.92	40.21	0.96

Tables S3 to S5 show the results of ARI, FMI, and AMI obtained by repeating

experiments (10 repetitions) of different methods on different datasets, corresponding to the performance of different methods in cell clustering (as shown in Figure. 2c-d of the main text).

Table S3. ARI of different methods in different datasets, related to Figure 2

datasets / methods	PBMC		HEART		LUAD		BC	
	mean	sd	mean	sd	mean	sd	mean	sd
GBC	0.16	0.03	0.23	0.05	0.26	0.05	0.03	0.01
plaid	0.16	0.04	0.01	0.02	0.01	0.04	0.00	0.00
Xmotif	0.04	0.02	0.00	0.00	0.00	0.00	0.00	0.00
CC	0.00	0.00	0.06	0.03	0.18	0.06	0.13	0.02
Bimax	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FABIA	0.04	0.01	0.01	0.01	0.05	0.01	0.00	0.00
autoCell	0.26	0.04	0.30	0.14	0.35	0.04	0.11	0.09
QUBIC2	0.00	0.00	0.00	0.00	0.06	0.03	0.15	0.11
DivBiclust	-0.00	0.00	-0.00	0.00				
scBC	0.30	0.03	0.36	0.07	0.46	0.09	0.19	0.09

Table S4. FMI of different methods in different datasets, related to Figure 2

datasets / methods	PBMC		HEART		LUAD		BC	
	mean	sd	mean	sd	mean	sd	mean	sd
GBC	0.33	0.04	0.37	0.04	0.48	0.05	0.33	0.02
plaid	0.28	0.03	0.04	0.11	0.07	0.17	0.00	0.00
Xmotif	0.25	0.04	0.00	0.00	0.00	0.00	0.00	0.00
CC	0.00	0.00	0.30	0.07	0.46	0.06	0.32	0.02
Bimax	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
FABIA	0.18	0.04	0.23	0.04	0.27	0.02	0.00	0.00
autoCell	0.38	0.03	0.41	0.16	0.51	0.02	0.29	0.16
QUBIC2	0.04	0.01	0.07	0.03	0.18	0.05	0.45	0.07
DivBiclust	0.11	0.01	0.11	0.01				
scBC	0.41	0.03	0.47	0.05	0.61	0.07	0.40	0.07

Table S5. AMI of different methods in different datasets, related to Figure 2

datasets / methods	PBMC		HEART		LUAD		BC	
	mean	sd	mean	sd	mean	sd	mean	sd
GBC	0.31	0.06	0.31	0.04	0.26	0.03	0.10	0.02
plaid	0.18	0.04	0.01	0.02	0.04	0.13	0.00	0.00
Xmotif	0.05	0.02	0.00	0.00	0.00	0.00	0.00	0.00
CC	0.00	0.00	0.06	0.03	0.12	0.05	0.08	0.01
Bimax	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FABIA	0.06	0.02	0.02	0.01	0.07	0.01	0.00	0.00
autoCell	0.44	0.04	0.42	0.16	0.47	0.04	0.18	0.11
QUBIC2	0.00	0.00	-0.00	0.01	0.06	0.03	0.21	0.11
DivBiclust	-0.00	0.01	-0.00	0.01				
scBC	0.47	0.03	0.47	0.05	0.53	0.08	0.25	0.07

Tables S6 to S11 show the results of 1-CE and F scores obtained by repeating experiments (100 repetitions for each setting) of different methods on simulated datasets under different settings. The values in parentheses represent standard deviations. They are corresponding to the comparison of the overall performance of biclustering (as shown in Figure. 3b-g of the main text) Bold items represent the best performance.

Table S6. 1-CE of different methods in different datasets (p=1000, n=300, L=3, $\times 10^{-3}$) , related to Figure 3

dropout / methods	0.2	0.3	0.4	0.5	0.55	0.6	0.65	0.7
GBC	6.21(1.07)	6.05(0.83)	5.31(1.47)	6.51(1.09)	5.71(0.80)	5.86(0.77)	6.00(1.01)	6.61(1.11)
plaid	12.96(2.07)	5.99(8.79)	1.64(4.61)	1.22(3.09)	0.68(2.11)	1.12(3.17)	0.92(2.49)	1.32(3.51)
Xmotif	0.01(0.05)	0.71(0.57)	1.03(0.95)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
CC	0.00(0.03)	0.01(0.05)	0.04(0.11)	0.14(0.18)	0.23(0.24)	0.30(0.32)	0.48(0.40)	0.57(0.45)
Bimax	2.02(1.10)	1.91(0.94)	1.70(0.68)	1.62(0.82)	1.63(0.88)	1.62(0.67)	1.40(0.88)	1.43(0.85)
FABIA	11.83(1.76)	11.57(1.64)	9.18(1.82)	6.95(1.02)	7.34(1.17)	7.25(1.25)	6.77(1.40)	5.81(1.13)
QUBIC2	0.32(0.31)	0.14(0.24)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
autoCell	9.86(1.76)	9.98(1.87)	9.50(2.07)	9.04(2.10)	9.51(2.09)	10.12(2.42)	9.98(2.47)	**9.92(2.59)**
scBC	**14.93(1.78)**	**14.83(1.60)**	**13.96(1.65)**	**13.98(1.78)**	**13.44(1.84)**	**12.24(2.15)**	**10.41(2.01)**	9.84(1.50)

Table S7. 1-CE of different methods in different datasets (p=3000, n=600, L=4, $\times 10^{-3}$), related to Figure 3

dropout / methods	0.2	0.3	0.4	0.5	0.55	0.6	0.65	0.7
GBC	2.87(0.69)	2.43(0.35)	2.41(0.34)	3.42(0.43)	2.98(0.36)	2.97(0.34)	2.95(0.33)	2.97(0.31)
plaid	**10.01(3.07)**	8.90(3.21)	1.84(3.16)	0.70(2.05)	0.65(1.65)	0.83(1.73)	0.68(1.66)	0.69(1.57)
Xmotif	0.01(0.05)	0.28(0.16)	0.00(0.03)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
CC	0.01(0.02)	0.01(0.02)	0.01(0.03)	0.05(0.05)	0.08(0.08)	0.11(0.10)	0.16(0.11)	0.22(0.15)
Bimax	0.60(0.24)	0.59(0.22)	0.46(0.22)	0.55(0.23)	0.52(0.23)	0.47(0.23)	0.50 (0.23)	0.45(0.25)
FABIA	9.29(1.07)	8.79(1.60)	6.91(1.55)	4.21(0.42)	5.41(0.78)	5.33(0.65)	5.13(0.70)	4.52(0.83)
QUBIC2	0.14(0.08)	0.04(0.04)	0.09(0.08)	0.07(0.06)	0.01(0.05)	0.03(0.08)	0.01(0.07)	0.01(0.06)
autoCell	7.65(1.02)	8.07(1.25)	8.19(1.03)	7.07(1.64)	6.25(1.52)	5.85(1.02)	4.82(1.33)	3.25(1.33)
scBC	9.39(1.04)	**9.89(1.00)**	**9.59(0.82)**	**9.45(0.95)**	**9.33(0.92)**	**8.67(1.20)**	**7.75(1.23)**	**6.36(1.13)**

Table S8. 1-CE of different methods in different datasets (p=6000, n=1500, L=5, $\times 10^{-3}$), related to Figure 3

dropout / methods	0.2	0.3	0.4	0.5	0.55	0.6	0.65	0.7
GBC	1.1(0.12)	1.16(0.09)	1.11(0.12)	1.45(0.15)	1.22(0.12)	1.22(0.12)	1.23(0.13)	1.21(0.12)
plaid	**5.21(1.18)**	3.1(1.5)	1.11(1.04)	0.69(1.49)	0.47(1.02)	0.54(1.13)	0.59(1.43)	0.16(0.55)
Xmotif	0.01(0.02)	0.11(0.07)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
CC	0(0)	0(0)	0.01(0.01)	0.01(0.02)	0.01(0.02)	0.01(0.01)	0.03(0.03)	0.06(0.04)
Bimax	0.32(0.1)	0.37(0.17)	0.31(0.19)	0.26(0.11)	0.27(0.1)	0.24(0.11)	0.23(0.11)	0.22(0.11)
FABIA	4.63(0.85)	3.94(0.76)	3.55(0.5)	2.39(0.26)	4.07(2.67)	3.31(1.4)	3.97(1.91)	3.4(1.79)
QUBIC2	0.06(0.05)	0.03(0.02)	0.03(0.03)	0.02(0.06)	0.02(0.05)	0.02(0.06)	0.02(0.06)	0.01(0.06)
autoCell	4.61(0.15)	4.75(0.17)	4.4(0.41)	4(0.46)	3.65(0.34)	3.53(0.32)	3.06(0.33)	2.96(0.45)
scBC	4.93(0.31)	**4.94(0.24)**	**4.81(0.36)**	**4.82(0.39)**	**4.8(0.38)**	**4.79(0.49)**	**4.7(0.47)**	**4.47(0.42)**

Table S9. F score of different methods in different datasets（p=1000, n=300, L=3，$\times 10^{-2}$）, related to Figure 3

dropout / methods	0.2	0.3	0.4	0.5	0.55	0.6	0.65	0.7
GBC	26.23(36.89)	6.93(6.53)	5.24(3.98)	13.08(3.87)	5.01(2.29)	6.30(5.84)	8.85(7.83)	10.32(6.5)
plaid	38.03(39.01)	11.31(16.30)	3.02(8.30)	2.26(5.72)	1.28(3.84)	2.04(5.90)	1.71(4.66)	2.19(5.62)
Xmotif	0.02(0.10)	1.44(1.14)	2.05(1.88)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)

CC	0.01(0.05)	0.02(0.10)	0.08(0.22)	0.30(0.38)	0.46(0.49)	0.62(0.68)	0.99(0.85)	1.16(0.95)
Bimax	3.97(2.22)	3.82(1.95)	3.39(1.49)	3.22(1.75)	3.28(1.80)	3.29(1.40)	2.83(1.80)	2.81(1.69)
FABIA	48.61(9.78)	47.15(9.30)	32.98(9.22)	26.21(8.17)	20.71(5.73)	17.64(4.79)	14.40(4.95)	10.15(2.5)
QUBIC2	0.65(0.63)	0.28(0.48)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
autoCell	16.23(12.34)	14.58(9.72)	13.56(7.68)	12.60(4.16)	13.53(5.09)	15.51(9.54)	9.98(4.47)	13.13(6.4)
scBC	**199.0(32.9)**	**177.7(56.7)**	**100.5(67.3)**	**50.57(48.32)**	**37.18(33.52)**	**33.19(34.09)**	**23.38(27.21)**	**15.39(5.4)**

Table S10. F score of different methods in different datasets（p=3000, n=600, L=4，$\times 10^{-2}$）, related to Figure 3

dropout / methods	0.2	0.3	0.4	0.5	0.55	0.6	0.65	0.7
GBC	2.51(2.73)	0.34(0.49)	0.17(0.25)	0.49(0.38)	0.16(0.03)	0.18(0.11)	0.17(0.07)	0.37(0.46)
plaid	2.99(1.51)	1.63(1.22)	0.34(0.57)	0.13(0.4)	0.12(0.3)	0.15(0.31)	0.12(0.29)	0.12(0.26)
Xmotif	0(0.01)	0.06(0.03)	0(0.01)	0(0)	0(0)	0(0)	0(0)	0(0)
CC	0(0)	0(0)	0(0.01)	0.01(0.01)	0.02(0.02)	0.02(0.02)	0.03(0.02)	0.04(0.03)
Bimax	0.12(0.05)	0.12(0.05)	0.09(0.05)	0.11(0.05)	0.1(0.05)	0.09(0.05)	0.1(0.05)	0.09(0.05)
FABIA	7.63(1.32)	4.77(0.99)	5.38(2.32)	2.66(0.55)	3.04(0.74)	2.49(0.58)	2.02(0.5)	**1.51(0.44)**
QUBIC2	0.03(0.02)	0.01(0.01)	0.02(0.02)	0.01(0.01)	0.01(0.01)	0.01(0)	0(0.01)	0(0)
autoCell	0.84(0.57)	0.96(0.48)	0.93(0.42)	0.86(0.37)	0.72(0.35)	0.68(0.3)	0.6(0.3)	0.54(0.21)
scBC	**8.45(1.79)**	**11.48(1.91)**	**11.46(1.61)**	**9.84(1.75)**	**8.41(3.12)**	**5.03(3.82)**	**3.03(3.06)**	1.1(1.01)

Table S11. F score of different methods in different datasets（p=6000, n=1500, L=5，$\times 10^{-2}$）, related to Figure 3

dropout / methods	0.2	0.3	0.4	0.5	0.55	0.6	0.65	0.7
GBC	0.59(0.89)	0.1(0.11)	0.08(0.05)	0.13(0.08)	0.05(0.04)	0.05(0.01)	0.06(0.04)	0.09(0.11)
plaid	2.09(2.25)	0.86(1.27)	0.21(0.22)	0.13(0.29)	0.09(0.19)	0.09(0.18)	0.11(0.26)	0.03(0.1)
Xmotif	0(0.01)	0.02(0.01)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
CC	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0.01(0.01)	0.01(0.01)
Bimax	0.06(0.02)	0.07(0.04)	0.06(0.04)	0.05(0.02)	0.05(0.02)	0.05(0.02)	0.05(0.02)	0.04(0.02)
FABIA	5.32(0.62)	4.9(0.7)	3.7(0.81)	2.72(0.51)	2.53(1.44)	2.15(1.24)	1.99(1.21)	1.46(0.89)
QUBIC2	0.01(0.01)	0.01(0)	0.01(0.01)	0(0.01)	0(0)	0(0)	0(0.01)	0(0)
autoCell	0.2(0.01)	0.26(0.07)	0.26(0.06)	0.25(0.07)	0.22(0.05)	0.22(0.06)	0.15(0.06)	0.11(0.05)
scBC	**5.76(0.37)**	**4.95(0.34)**	**3.89(0.65)**	**2.91(0.65)**	**2.82(0.63)**	**2.71(0.7)**	**2.54(0.72)**	**1.74(0.78)**